School of Computing and Information Systems,
The University of Melbourne

Machine Learning for Feedback in Massive Open Online Courses

Jiazhen He

Submitted in total fulfilment of the requirements
of the degree of Doctor of Philosophy

Produced on archival quality paper

December, 2016

Abstract

Massive Open Online Courses (MOOCs) have received widespread attention for their potential to scale higher education, with multiple platforms such as Coursera, edX and Udacity recently appearing. Online courses from elite universities around the world are offered for free, so that anyone with internet access can learn anywhere. Enormous enrolments and diversity of students have been widely observed in MOOCs. Despite their popularity, MOOCs are limited in reaching their full potential by a number of issues. One of the major problems is the notoriously low completion rates. A number of studies have focused on identifying the factors leading to this problem. One of the factors is the lack of interactivity and support. There is broad agreement in the literature that interaction and communication play an important role in improving student learning. It has been indicated that interaction in MOOCs helps students ease their feelings of isolation and frustration, develop their own knowledge, and improve learning experience. A natural way of improving interactivity is providing feedback to students on their progress and problems.

MOOCs give rise to vast amounts of student engagement data, bringing opportunities to gain insights into student learning and provide feedback. This thesis focuses on applying and designing new machine learning algorithms to assist instructors in providing student feedback. In particular, we investigate three main themes: i) identifying at-risk students not completing courses as a step towards timely intervention; ii) exploring the suitability of using automatically discovered forum topics as instruments for modelling students' ability; iii) similarity search in heterogeneous information networks. The first theme can be helpful for assisting instructors to design interventions for at-risk students to improve retention. The second theme is inspired by recent research on mea-

surement of student learning in education research communities. Educators explore the suitability of using latent complex patterns of engagement instead of traditional visible assessment tools (e.g. quizzes and assignments), to measure a hypothesised distinctive and complex learning skill of promoting learning in MOOCs. This process is often human-intensive and time-consuming. Inspired by this research, together with the importance of MOOC discussion forums for understanding student learning and providing feedback, we investigate whether students' participation across forum discussion topics can indicate their academic ability. The third theme is a generic study of utilising the rich semantic information in heterogeneous information networks to help find similar objects. MOOCs contain diverse and complex student engagement data, which is a typical example of heterogeneous information networks, and so could benefit from this study.

We make the following contributions for solving the above problems. Firstly, we propose transfer learning algorithms based on regularised logistic regression, to identify students who are at risk of not completing courses weekly. Predicted probabilities with well-calibrated and smoothed properties can not only be used for the identification of at-risk students but also for subsequent interventions. We envision an intervention that presents probability of success/failure to borderline students with the hypothesis that they can be motivated by being classified as "nearly there". Secondly, we combine topic models with measurement models to discover topics from students' online forum postings. The topics are enforced to fit measurement models as statistical evidence of instruments for measuring student ability. In particular, we focus on two measurement models, the Guttman scale and the Rasch model. To the best our knowledge, this is the first study to explore the suitability of using discovered topics from MOOC forum content as instruments for measuring student ability, by combining topic models with psychometric measurement models in this way. Furthermore, these scaled topics imply a range of difficulty levels, which can be useful for monitoring the health of a course and refining curricula, student assessment, and providing personalised feedback based on student ability levels and topic difficulty levels. Thirdly, we extend an existing meta path-based similarity measure by incorporating transitive similarity and tem-

poral dynamics in heterogeneous information networks, evaluated using the DBLP bibliographic network. The proposed similarity measure might apply to MOOC settings to find similar students or threads, or thread recommendation in MOOC forums, by modelling student interactions in MOOC forums as a heterogeneous information network.

Declaration

This is to certify that

1. the thesis comprises only my original work towards the degree of Doctor of Philosophy except where indicated in the Preface,
2. due acknowledgement has been made in the text to all other material used,
3. the thesis is fewer than 80,000 words in length, exclusive of tables, maps, bibliographies and appendices.

Jiazhen He

Preface

This thesis has been written at the Department of Computing and Information Systems, The University of Melbourne. The main contents of the thesis are the Chapters 3, 4, 5 and 6. They are based on manuscripts published or under review for publication. I declare that I am the primary author and have contributed $> 50\%$ in the following papers.

- Jiazhen He, James Bailey, Benjamin I.P. Rubinstein, and Rui Zhang. "Identifying At-Risk Students in Massive Open Online Courses". In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1749–1755. AAAI Press, 2015. (Chapter 3 contains material from this paper.)

- Jiazhen He, Benjamin I.P. Rubinstein, James Bailey, Rui Zhang, Sandra Milligan, and Jeffrey Chan. "MOOCs Meet Measurement Theory: A Topic-Modelling Approach". In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1195–1201. AAAI Press, 2016. (Chapter 4 contains material from this paper.)

- Jiazhen He, Rui Zhang, James Bailey, Benjamin I.P. Rubinstein, and Sandra Milligan. "TopicResponse: A Marriage of Topic Modelling and Rasch Modelling for Automatic Measurement in MOOCs". Under second round major revision for *Machine Learning Journal*, 2016. (Chapter 5 contains material from this paper.)

- Jiazhen He, James Bailey, and Rui Zhang. "Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks". In *International Conference on Database Systems for*

Advanced Applications, pages 141–155. Springer, 2014. (Chapter 6 contains material from this paper.)

Acknowledgments

PhD study is a tough but amazing journey of learning and exploration, with ups and downs just like Melbourne's random weather. I would not be able to complete this thesis, the significant milestone of my PhD journey, without the help and support from many people.

First of all, I would like to express my sincere gratitude to my supervisors: Prof. Rui Zhang, Prof. James Bailey, and Dr. Benjamin Rubinstein for their continuous encouragement and guidance for my PhD study. Thanks Rui for accepting me as a PhD student, and giving me the freedom to choose research topics. He has been very supportive and provided many helpful suggestions. His high standard for himself always inspire me to work hard. Thanks James for his kindness and patience for clarifying my doubts about my PhD study and research. He was always there with a smiling face whenever I walked into his office for advice, and provided me very insightful suggestions. His positive attitude has been the important source to keep me motivated and persist. Thanks Ben for his very detailed technical discussions and comments which helped me think and understand deeply. His expertise in machine learning and open mind have inspired me to think in a different way. At the same time, he is a careful person, helping me polish paper like an art. He has been always actively providing me helpful advices and suggestions. I thank him for his connection with IBM research, providing me an internship opportunity. I feel so lucky to work with three wonderful supervisors during my PhD study.

I am also thankful to my collaborators Jeffrey Chan for his insightful comments and discussions, helping me better understand algorithms. Thanks go to Sandra Milligan for her interest on our work and the opportunity for collaboration. Her work on measurement has been an important sources of inspiration

for our work. I would also like to thank the Learning Analytic Research Group lead by Prof. Gregor Kennedy for providing such a good opportunity to expose people and research in the education and psychology community. I am also grateful to Aaron Howard for serving as my committee member, supporting the plan of my research at all milestones and providing useful comments.

I would like to thank The University of Melbourne, the Computing and Information Systems (CIS) Department, and the Head of the Department Justin Zobel for providing such good conditions where I have been exposed to great minds. I am also grateful to Data61/NICTA, for supporting me for the living and travelling expenses during my PhD study. I would also like to thank IBM Research Australia for the great internship experience. Special thanks to Wanita Sherchan, Jey Han Lau, Christopher Butler and Mukesh Mohania.

I am very grateful to do my research in a vibrant office. I would like to thank all my colleagues in my office. We shared research experience and life experience with research discussions and casual talks. Special thanks to Simone Romano for his kindness and insightful discussions, Sergey Demyanov and Goce Ristanoski for all the great time, laughter and fun we shared. I would also like to thank all my other colleagues in my office and from other offices of the CIS departments: Yamuna, Yang, Florin, Yun, Shuo, Alvin, Daniel, Mohadeseh, Donia, Jianzhong, Jin, Xiaojie, Zeyi, Yu, Andy, Yuan, Saad, Gitansh, Yiqing, Han, Chenghao, Bowen, Safiollah, Zay, and many others. I am also grateful to my friends: Liyan and Yanhui for supporting and encouraging me to go through the tough days in research and personal life.

Last but not the least, I would like to thank my family for their spiritual support, constant support and unconditional love.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Massive Open Online Courses (MOOCs) have gained tremendous popularity since 2012, "The Year of the MOOC" (Pappano, 2012), when multiple platforms such as Coursera, edX and Udacity were launched. MOOCs aim to make higher education accessible to the world, by offering online courses from a range of elite universities for free. Enormous enrolments have been widely enjoyed in MOOCs — the average course enrolment has been found around to be 43,000 students (Jordan, 2014). For example, according to Hare (2016), The University of Melbourne has surpassed its millionth MOOC enrolment, with 20 MOOCs offered on Coursera. Beyond the large scale of enrolments, MOOCs have attracted a diverse population of students [1] from a variety of age groups, educational backgrounds and nationalities. MOOCs are becoming more mainstream, as they continue to grow in size and diversity in terms of both students and courses.

Despite their popularity, MOOCs are limited in reaching their full potential by a number of problems. One of the major problems is the notoriously low completion rates — the average completion rate has been found to be below 7% (Jordan, 2014). To take an example, Table 1.1 shows the student participation in the first offering of a Coursera MOOC *Discrete Optimisation* lauched by The University of Melbourne in 2013. Of 51,306 students enrolled, only 795 students completed: a completion rate of just 1.5%. And only 27,679 (about 54%)

[1] Interchangeable with learners, participants.

Table 1.1: Student participation in the first offering of *Discrete Optimisation* launched in 2013; actions are measured in terms of viewing/downloading lectures and completing quizzes/assignments.

	Discrete Optimisation MOOC
Number of students enrolled	51,306
Number of students with actions	27,679
Number of students completed	795

students ever engaged in lectures and quizzes/assignments; even restricted to this group, the completion rate was a mere 2.9%. Although it is arguable to use completion rates to evaluate MOOC success (Rivard, 2013), it is important to study the factors leading to dropout, in order to improve learning experience and overall satisfaction. A large body of work has studied the reasons for the low completion rates, with most significant factors suggested as: no intention to complete (Kolowich, 2013; Onah et al., 2014), lack of time (Onah et al., 2014; Khalil and Ebner, 2014), lack of interactivity and support in MOOCs (Khalil and Ebner, 2014; Mackness et al., 2010; Onah et al., 2014), and insufficient background and skills (Khalil and Ebner, 2014). There is broad agreement in the literature that interaction and communication play an important role in improving student learning. It has been indicated that interaction in MOOCs helps students ease their feelings of isolation and frustration, develop their own knowledge, and improve learning experience (Khalil and Ebner, 2014). A natural way of improving interactivity is to provide feedback to students on their progress and problems.

The scale and diversity of MOOC students, however, create challenges for providing feedback. Unlike traditional small classes where students can receive immediate feedback from instructors, it is impossible for an instructor to interact with the huge cohort of students enrolled in MOOCs. Additionally, diverse student backgrounds and motivations call for personalised feedback that is adapted to their needs and goals.

1.1 Research Questions

MOOCs produce vast amounts of student engagement data, bringing new opportunities to gain insights into student learning and provide feedback, to improve learning experience and outcomes. This thesis focuses on applying and designing new machine learning algorithms to assist instructors in providing student feedback. In particular, we investigate the following three research themes:

1. **Identifying at-risk students**: Can we identify at-risk students not completing courses accurately and early with a view to timely intervention? What feedback can we provide to help them?

2. **Topic-instrumented measurement**: What topics do students discuss in MOOC discussion forums? Can we discover topics such that students' participation in them can be used for modelling their ability?

3. **Similarity search in heterogeneous information networks**: Can the rich semantic information in heterogeneous information network help improve similarity search?

The first theme can be useful in helping instructors design timely interventions to improve retention. In addition to the identification of at-risk students, we suggest presenting the probability of success/failure to borderline students as an intervention. This creates challenges for the predicted probabilities: they are required to be well-calibrated and smoothed across weeks.

The second theme is inspired by the importance of forum discussions for understanding student learning, and recent research on quantitative measurement of student learning in the education community. In particular, i) MOOC discussion forums, as the main platform for student-instructor and student-student interactions, is of importance in gaining insights into student learning. ii) Recent research in education (Milligan, 2015) suggests that a distinctive and complex learning skill is required to generate learning in MOOCs. Educators are interested in whether and how the possession of this skill may be evidenced by latent complex patterns of engagement, instead of traditional assessment tools

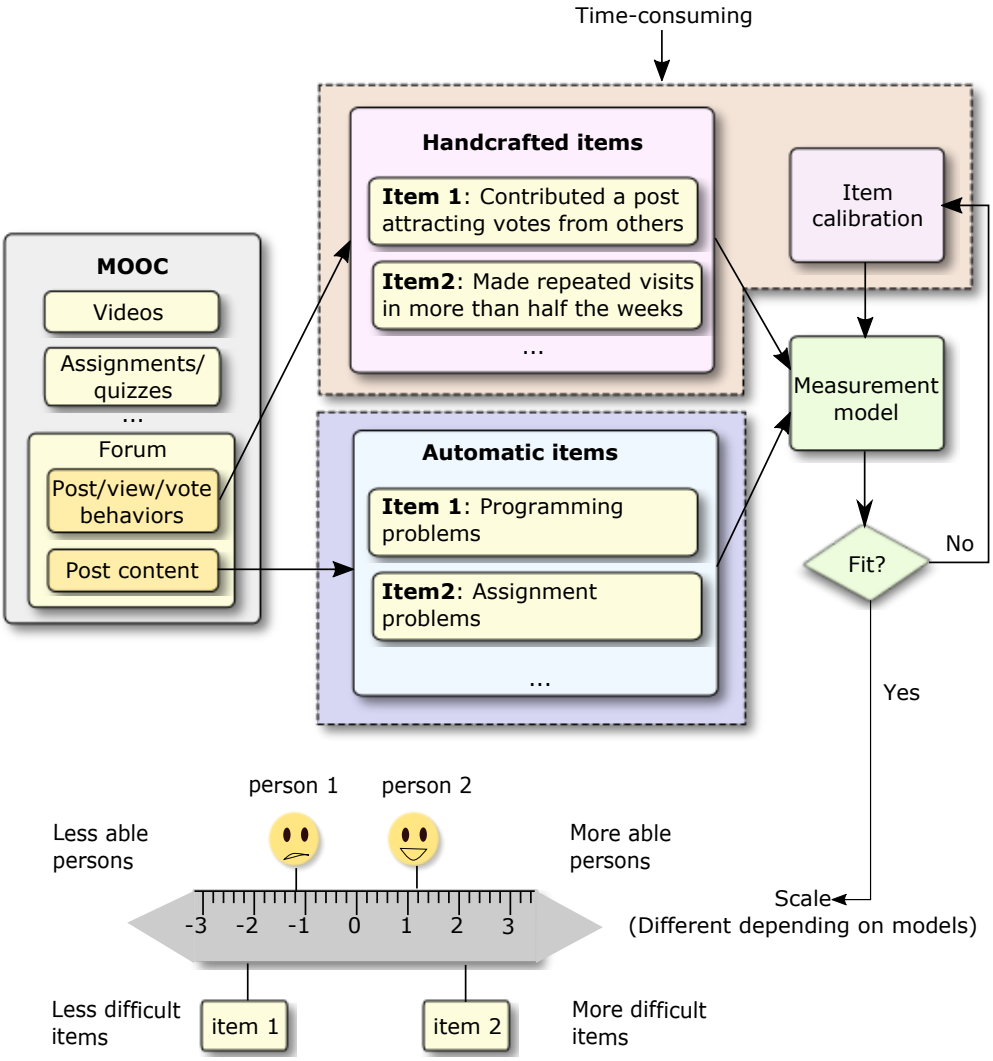


Figure 1.1: Workflow for devising items manually versus automatically discovering topics as items for measurement. Traditionally, a set of items are handcrafted from MOOC forum behaviours, and then the students' responses on the items are examined using a measurement model. If the model fits well, then the students and items can be compared on an inferred scale, depicted as a ruler. Otherwise, the items are refined (changed, added or deleted) manually until model fit. The process of handcrafting items and calibration is time-consuming. Instead, we aim to automatically generate topics from discussion posts as items that fit a measurement model by design.

such as quizzes and assignments. iii) In order to validate such a hypothesis, measurement theory can be used. A set of items is handcrafted from forum

activities (*e.g.*, "*contributed a post attracting votes from others*" and "*made repeated thread visits in more than half the weeks*"), and calibrated (*e.g.*, deleted or changed) to fit a measurement model (Milligan, 2015). This process is human-intensive and time-consuming as shown in Figure 1.1. Driven by the above observations, we investigate whether students' participation in automatically discovered forum discussion topics can be used as an instrument to model students' ability. To establish the statistical effectiveness, students' participation in topics are required to fit a measurement model as evidence of reliability as shown in Figure 1.1. The resulting scaled topics can show difficulty levels produced from measurement models, and could be helpful for curriculum refinement, student assessment, and personalised adaptive feedback. The challenge is to automatically discover topics such that students' participation in them fit a measurement model.

Thirdly, we study a generic problem of similarity search in heterogeneous information networks. It is inspired by the rich, diverse and complex student engagement information in MOOCs (*e.g.*, engagement in videos, quizzes and forum discussions), framing a MOOC as a typical heterogeneous information network. We extend an existing meta path-based similarity measure by incorporating transitive similarity and temporal dynamics, evaluated in a DBLP bibliographic network domain. The challenge is to model the transitive similarity and temporal dynamics for similarity search. Our approach could be applied to MOOC settings, *e.g.*, for finding similar students or threads, or thread recommendation, by modelling student interactions in MOOC forums as a heterogeneous information network.

1.2 Thesis Overview

This thesis contributes to algorithms for solving the above problems with the goal of providing feedback in MOOCs.

In Chapter 2, we introduce background on machine learning research for MOOCs, particularly on identifying at-risk students and MOOC forums. Then we give a brief overview of the topic modelling approach of non-negative ma-

trix factorisation, and topic modelling research for MOOCs. Next, we review background in measurement and item response theory with a focus on two measurement models we used in thesis, the Guttman scale and the Rasch model. Finally, we review research on measurement models used in machine learning and MOOCs.

In Chapter 3, we focus on the first theme — accurate and early identification of students who are at risk of not completing courses. We build predictive models weekly, over multiple offerings of a course. Furthermore, we envision an intervention that presents meaningful pass/failure probability to borderline students, with the hypothesis that they could be motivated by being "nearly there". To be effective, predicted probabilities are required to be both well-calibrated and smoothed across weeks. Based on logistic regression, we propose two transfer learning algorithms to trade-off smoothness and accuracy, by adding a regularisation term to minimise the difference of success/failure probabilities between consecutive weeks. Experimental results on two offerings of a Coursera MOOC establish the effectiveness of our algorithms.

In Chapter 4, we focus on the second theme, and consider a simple measurement scale —the Guttman scale (Guttman, 1950), for modelling students' ability. In particular, we adapt the topic modelling approach of Non-Negative Matrix Factorisation (NMF) to discover topics from students' online forum postings, such that students' participation across these topics conforms to the Guttman scale, as statistical evidence of reliability. This is done by introducing a novel regularisation into NMF to favour Guttman-scaled topics. Students and topics can be ordered and used for curriculum refinement and student assessment. We demonstrate the suitability of our approach with both quantitative experiments on three Coursera MOOCs, and with a qualitative survey of topic interpretability on two MOOCs by domain expert interviews.

In Chapter 5, we also focus on the second theme but consider a more widely used measurement model, the Rasch model (Rasch, 1960) from item response theory. The Rasch model, stochastic extension of the Guttman scale, has several advantages. It allows interval measurement, while the Guttman scale only allows ordinal measurement. This makes it possible to provide personalised feedback and adaptive testing. To establish the statistical effectiveness of us-

ing topics as instruments to model student ability, we combine topic modelling with Rasch modelling. The discovered Rasch-scaled topics may assist instructors in curriculum refinement, student assessment, and providing personalised feedback. We provide a quantitative validation on three Coursera MOOCs, and a qualitative examination of topic interpretation with inferred difficulty levels on a Discrete Optimisation MOOC.

In Chapter 6, we study the third theme of finding similar objects in heterogeneous information networks, where rich semantic information are available. We extend an meta path-based similarity measure by incorporating richer information, such as transitive similarity and temporal dynamics. Experiments on a large DBLP network show that our improved similarity measure is more effective at identifying similar authors in terms of their future collaborations. The proposed similarity measure could apply to MOOC settings.

Finally, Chapter 7 concludes this thesis, and discusses future works.

Chapter 2

Background

This chapter begins by introducing machine learning research for MOOCs, with a focus on identifying at-risk students and MOOC forums. Then we introduce topic modelling with an emphasis on non-negative matrix factorisation, and its applications for MOOCs. Next, we review background in measurement and item response theory, with a focus on the two measurement models used in this thesis, the Guttman scale and the Rasch model. Lastly, we introduce research on measurement used in MOOCs and machine learning. This chapter focuses on the related work on the common theme of MOOCs. The related work on similarity search in heterogeneous networks is discussed in Chapter 6.

2.1 Machine Learning Research for MOOCs

The booming popularity of MOOCs, has attracted widespread attention from researchers in the computer science community. A large number of studies on MOOCs have recently appeared in both education and data mining related venues, *e.g.*, the International Conference on Educational Data Mining (EDM) and the International Conference on Artificial Intelligence in Education (AIED) which has a long history of interdisciplinary research on artificial intelligence, education and psychology. MOOCs are also a popular topic of young conferences such as the International Learning Analytics and Knowledge Conference (LAK) and the ACM Conference on Learning@Scale (L@S). Furthermore, work-

shops in conjunction with traditional machine learning and natural language processing conferences also showcase research contributions on MOOCs, *e.g.*, NIPS 2013 Workshop on Data Driven Education, EMNLP 2014 Workshop on Modelling Large Scale Social Interaction in Massively Open Online Courses, ICML 2015 Workshop on Machine Learning for Education, and NIPS 2016 Workshop on Machine Learning for Education.

Various studies have been conducted into MOOCs for tasks such as dropout prediction (Halawa et al., 2014a; Yang et al., 2013; Ramesh et al., 2014b; Kloft et al., 2014; He et al., 2015), characterising student engagement (Anderson et al., 2014; Kizilcec et al., 2013; Ramesh et al., 2014b) and peer assessment (Dıez et al., 2013; Piech et al., 2013; Mi and Yeung, 2015). We focus on introducing related works for identifying at-risk students and for analysing MOOC forums.

2.1.1 Identifying At-Risk Students

Low completion rates have been a criticism made about MOOCs right from their beginnings. There has been a large amount of work focusing on success/failure or dropout prediction. Ramesh et al. (2013) analyse students' online behavior and identify two types of engagement, which is then used as a latent feature to help predict final performance. The same methodology is then used to predict a similar task for whether students submitted their final quizzes/assignments (Ramesh et al., 2014c). However, these predictions were not studied with a view to intervention. Instead, we propose to intervene students by presenting meaningful predicted probabilities, with only those who are on the pass/fail borderline targeted. Stages of targeted interventions have parallels to epidemiological approaches to education (Lodge, 2011) and are conceptually similar to defence-in-depth and perimeter defences in computer security (Kaufman et al., 2002).

A similar task is dropout prediction, has class label whether or not a student will dropout instead of fail. While we have not focused on dropout prediction, our techniques could readily apply to this setting. Most studies focus on developing features from students' behaviours and engagement patterns to help prediction. Kloft et al. (2014) predict dropout from only click-stream data us-

ing a Support Vector Machine (SVM). Taylor et al. (2014) utilise crowd-sourced feature engineering (Veeramachaneni et al., 2014) to predict dropout based on logistic regression. Balakrishnan (2013) extracts features mainly from discussion forums and video lectures, and employs Hidden Markov Models (HMMs) to predict student retention. Halawa et al. (2014b) study accurate and early dropout prediction using student activity features capturing lack of ability or interest.

Previous work has concentrated on using different data sources, carrying out feature engineering and using off-the-shelf classifiers evaluated only within one offering of a course. However to the best of our knowledge, none have i) recognised the importance of calibrated prediction probabilities for predicting failure or dropout; ii) explored and motivated the need for temporally smooth prediction probabilities in the context of education and interventions; iii) applied transfer learning for this purpose; and iv) shown that a model trained a previous MOOC offering can be used effectively for predicting within a future offering.

Another research area exploring low completion rates is correlative analysis to understand factors influencing success/failure or dropout/retention. Various factors have been investigated, such as demographics (DeBoer et al., 2013b,a), student behavior and social positioning in forums (Yang et al., 2013), sentiment in forums (Wen et al., 2014) and peer influence (Yang et al., 2014b). This can help better understand the reasons for success/failure or dropout/retention and potentially help devise detailed feedback, but it is not our focus in this thesis.

2.1.2 Machine Learning for Analysing MOOC Forums

MOOC forums have been of great interest recently, due to the availability of rich textual data and social behaviour. Various studies have been conducted, such as sentiment analysis, community finding, question recommendation, answers predication and intervention prediction. Wen et al. (2014) use sentiment analysis to monitor students' trending opinions towards the course and to correlate sentiment with dropouts over time using survival analysis. Yang et al. (2015)

predict students' confusion with learning activities, as expressed in the discussion forums using discussion behaviour and clickstream data, and explore the impact of confusion on student dropout. Ramesh et al. (2015) predict sentiment in MOOC forums using hinge-loss Markov random fields. Yang et al. (2014a) study question recommendation in discussion forums based on matrix factorisation. Gillani et al. (2014) find communities using Bayesian Non-Negative Matrix Factorisation. Yang et al. (2014a) recommend questions of interest to students by designing a context-aware matrix factorisation model considering constraints on students and questions. MOOC forum data has also been studied for the task of predicting the accepted answer to a forum question (Jenders et al., 2016) and predicting the instructor intervention (Chaturvedi et al., 2014). Despite the variety of studies, little machine learning research has explored forum discussions for the purpose of measurement in MOOCs.

2.2 Topic Modelling

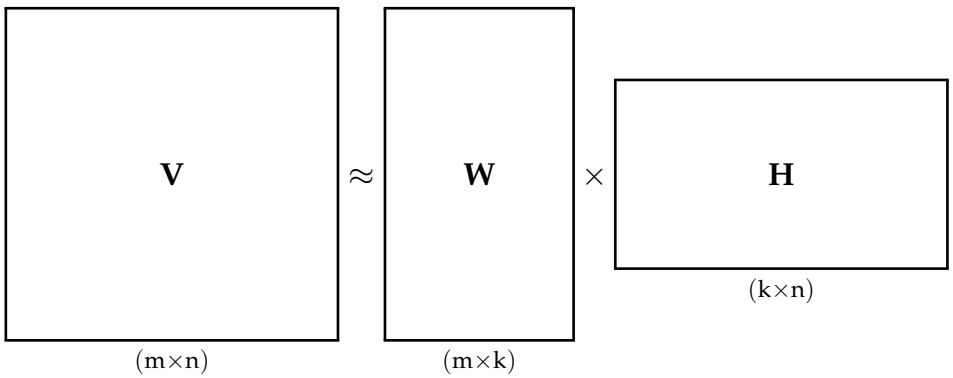
Topic modelling is a powerful tool for analysis of text data, and has been widely used in the machine learning community to identify latent topic structure in a corpus of documents. A topic is often represented as a mixture of words, and a document is represented as a mixture of topics. Given a corpus of documents, topic modelling provides methods to finding the assignments of words to topics, and the assignments of topics to documents. Latent Dirichlet Allocation (LDA) (Blei et al., 2003) and Non-Negative Matrix Factorisation (NMF) (Lee and Seung, 1999) are two popular techniques for topic modelling. In this thesis, we focus on NMF, due to the interpretability of topics produced, and the feasibility of extending its optimisation objective to admit natural integration with education constraints.

2.2.1 Non-Negative Matrix Factorisation (NMF)

Given a non-negative matrix $\mathbf{V} = (v_{ij}) \in \mathbb{R}^{m \times n}$ and a positive integer k, NMF factorises $\mathbf{V}$ into the product of a non-negative matrix $\mathbf{W} = (w_{ij}) \in \mathbb{R}^{m \times k}$ and a

non-negative matrix $\mathbf{H} = (h_{ij}) \in \mathbb{R}^{k \times n}$

$$\mathbf{V} \approx \mathbf{WH}$$



A commonly-used measure for quantifying the quality of this approximation is the Frobenius norm between $\mathbf{V}$ and $\mathbf{WH}$. Thus, NMF involves solving the following optimisation problem,

$$\min_{\mathbf{W},\mathbf{H}} \|\mathbf{V} - \mathbf{WH}\|_F^2 \quad \text{s.t.} \quad \mathbf{W} \geq \mathbf{0}, \;\; \mathbf{H} \geq \mathbf{0} \;\; . \tag{2.1}$$

The objective function is convex in $\mathbf{W}$ and $\mathbf{H}$ separately, but not together. Therefore standard optimisers are not expected to find a global optimum. The multiplicative update algorithm (Lee and Seung, 2001) is commonly used to find a local optimum, where $\mathbf{W}$ and $\mathbf{H}$ are updated by a multiplicative factor that depends on the quality of the approximation.

$$\begin{aligned} w_{ij} &\leftarrow w_{ij} \frac{(\mathbf{VH}^T)_{ij}}{(\mathbf{WHH})^T} \\ h_{ij} &\leftarrow h_{ij} \frac{(\mathbf{W}^T\mathbf{V})_{ij}}{(\mathbf{W}^T\mathbf{WH})_{ij}} \end{aligned}$$

In the present MOOC setting, we focus on the students who contributed posts or comments in forums. For each student, we aggregate all posts or comments that they contributed. Each student is represented by a bag of words as shown in the example word-student matrix $\mathbf{V}$ in Figure 2.1, where m represents the number of words, and n represents the number of students. Using NMF, a

$$\mathbf{V} = \begin{array}{c} \\ \text{solver} \\ \text{optim} \\ \text{code} \\ \text{algorithm} \\ \vdots \\ \text{word}m \end{array} \begin{array}{c} \begin{array}{cccc} \text{stud1} & \text{stud2} & \cdots & \text{stud}n \end{array} \\ \begin{pmatrix} 0.26 & 0.11 & \cdots & 0.52 \\ 0.32 & 0.18 & \cdots & 0.06 \\ 0.68 & 0.01 & \cdots & 0.83 \\ 0.89 & 0.61 & \cdots & 0.44 \\ \vdots & \vdots & \ddots & \vdots \\ 0.22 & 0.54 & \cdots & 0.98 \end{pmatrix} \end{array}$$

$$\mathbf{W} = \begin{array}{c} \\ \text{solver} \\ \text{optim} \\ \text{code} \\ \text{algorithm} \\ \vdots \\ \text{word}m \end{array} \begin{array}{c} \begin{array}{cccc} \text{topic1} & \text{topic2} & \cdots & \text{topic}k \end{array} \\ \begin{pmatrix} 0.22 & 0.01 & \cdots & 0.12 \\ 0.38 & 0.15 & \cdots & 0.06 \\ 0.18 & 0.05 & \cdots & 0.03 \\ 0.09 & 0.21 & \cdots & 0.01 \\ \vdots & \vdots & \ddots & \vdots \\ 0.02 & 0.04 & \cdots & 0.12 \end{pmatrix} \end{array}$$

$$\mathbf{H} = \begin{array}{c} \\ \text{topic1} \\ \text{topic2} \\ \vdots \\ \text{topic}k \end{array} \begin{array}{c} \begin{array}{cccc} \text{stud1} & \text{stud2} & \cdots & \text{stud}k \end{array} \\ \begin{pmatrix} 0.83 & 0.17 & \cdots & 0.04 \\ 0.21 & 0.75 & \cdots & 0.16 \\ \vdots & \vdots & \ddots & \vdots \\ 0.09 & 0.64 & \cdots & 0.62 \end{pmatrix} \end{array}$$

Figure 2.1: Example matrices: word-student **V**, word-topic **W**, topic-student **H**.

word-student matrix **V** can be factorised into two non-negative matrices: word-topic matrix **W** and topic-student matrix **H**. For each student, the column vector of **V** is approximated by a linear combination of the columns of **W**, weighted by the components of **H**. Therefore, each column vector of **W** can be regarded as a topic, and the memberships of students in these topics are encoded by **H** as shown in Figure 2.1.

2.2.2 Topic Modelling for MOOCs

Topic models and their variants have been used for a wide variety of applications, such as information retrieval to find relevant document to a query (Wei and Croft, 2006), modelling author influences (Gerrish and Blei, 2010), modelling citation influences (Dietz et al., 2007), word sense disambiguation (Boyd-Graber et al., 2007) and part-of-speech tagging (Toutanova and Johnson, 2007). Beyond the applications for text data, topic models have also been used for understanding images (Fei-Fei and Perona, 2005; Chong et al., 2009), biology

(Pritchard et al., 2000) and music (Hu and Saul, 2009).

In MOOCs, topic modelling has been applied for tasks such as understanding key themes in forum discussions (Robinson, 2015; Atapattu and Falkner, 2016), predicting student survival (Ramesh et al., 2014a), study partner recommendation (Xu and Yang, 2015) and course recommendation (Apaza et al., 2014). Apart from analysing the content information from forum discussions or course syllabi, topic models have been adapted to analyse student behaviours from clickstream data(Wen and Rosé, 2014; Coleman et al., 2015). However, to our knowledge, no studies have leveraged topics generated from topic modelling as instruments for measurement, and algorithmically combined topic models with measurement models from psychometrics.

2.3 Measurement

Measurement in education and psychology is the process of assigning a number to an attribute of an individual according to a rule that individuals can be compared to one another (Pedhazur and Schmelkin, 1991). Unlike measurement in the physical world, where attributes like height, weight and age are observable, the attributes to be measured in education and psychology are often intangible, such as attitudes, abilities or intelligence. Qualitative information must often be converted into a numerical form for further analysis. Measurement and scaling techniques make this conversion possible.

Since the attribute to be measured is not directly observable, a set of items is often devised manually and individuals' responses on the items are collected. Based on a modelled correspondence with observed item responses, latent attribute levels of a cohort can be inferred. This process is typically called *scaling* (De Ayala, 2013). It involves the methods of developing a scale. In the following, we firstly introduce the levels of measurement, and then introduce a measurement scale, the Guttman scale, which is used in this thesis.

2.3.1 Levels of Measurement

The level of measurement refers to the relationship among the numbers that are assigned to the measured attributes of an individual (or an object in a broad sense). Each level possesses different properties, allowing different interpretations about the assigned number and different statistical analysis on them. There are typically four levels of measurement scales (from low to high). Table 2.1 shows the summary of characteristics and examples for levels of measurement.

- Nominal scale
  This is the lowest level of measurement, often referred to as a categorical scale. It classifies individuals into categories, and no ordering of individuals is implied. The numbers are simply used as identifiers to represent categories of individuals. For example, gender can be measured on a nominal scale, with two categories: male and female (may be coded as 1 and 2 respectively).

- Ordinal scale
  This scale allows ordering or ranking of objects. It provides information about direction or order, in addition to the nominal information provided in a nominal scale. For example, the first, second and third place in a race.

- Interval scale
  With an interval scale, we know not only the order, but also the exact differences between the values. It possesses equal intervals. For example, measuring temperature in Celsius is an example of an interval scale. The difference between 40 and 30 degrees is the same as the difference between 20 and 10 degrees. But we cannot say 40 degrees is twice as hot as 20 degrees, which is a property of a ratio scale.

- Ratio scale
  This is the highest level of a measurement scale. In addition to the properties in an interval scale, it has an absolute zero, which allows meaningful interpretation of ratios. For example, time is a ratio scale, with meaningful ratio interpretation: *e.g.*, 30 minutes is twice as long as 15 minutes.

Table 2.1: Summary of characteristics and examples for levels of measurement.

Scale	Characteristics	Examples
Nominal	Categories; No ordering or ranking	Gender; Marital status
Ordinal	Ranking, order	Ranking in a race, class
Interval	Meaningful difference/distance between scale values but no true zero (no meaning about ratio)	Temperature
Ratio	True zero exists, ratios of scale values can be compared	Length, weight, height, age, time

2.3.2 The Guttman Scale

Guttman scaling (Guttman, 1950) was developed in the 1940s, and is used in social psychology and education. The goal of Guttman scaling is to establish unidimensional measurement instruments. It induces a total ordering on items—an individual who successfully answers/agrees with a particular item also answers/agrees with items of lower rank-order. Table 2.2 depicts an example Guttman scale measuring mathematical ability (Abdi, 2010), where the items are ordered in increasing latent difficulty, from *Counting* to *Division*. Here the total score corresponds to the persons' latent ability: the greater the higher.

Table 2.2: An example of a perfect Guttman scale measuring mathematical ability (Abdi, 2010) , where 1 means the person has mastered the item and 0 for not. Person 5 who has mastered the most difficult item *Division*, is expected to have mastered all easier items as well.

	Item 1 (Counting)	**Item 2 (+)**	**Item 3 (−)**	**Item 4 (×)**	**Item 5 (÷)**	**Total Score**
Person 1	1	0	0	0	0	1
Person 2	1	1	0	0	0	2
Person 3	1	1	1	0	0	3
Person 4	1	1	1	1	0	4
Person 5	1	1	1	1	1	5

If a set of items forms a Guttman scale, persons' responses on items should be predicted from their total score, which is simply the sum of the items they

answer correctly or agree with. For example, by knowing person 4's total score which is 4, the response pattern on the set of items 11110 can be reproduced. As it is rare to construct a perfect Guttman scale in practice, metrics are needed to evaluate Guttman scale quality. The coefficient of reproducibility (CR) is often used to measure reproducibility, which is equal to the proportion of correct predictions. A scale is often revised by changing, adding or removing items until it can produce an acceptable CR, 0.9 by convention (Torgerson, 1958).

$$CR = 1 - \frac{\text{No. of errors}}{\text{No. of possible errors(Total responses)}} \ .$$

The Guttman scale provides ordinal measurement, which only allows ranking of persons, but not comparing the attribute difference between persons. For example, in Table 2.1, person 4 is more able than person 3, but there is no information about how much person 4 is better than person 3.

2.4 Item Response Theory (IRT)

IRT (Lord, 1980), also known as latent trait theory, studies statistical models for measurement in education and psychology. Such models specify the probability of an individual's response on an item as a mathematical function of the individual's and item's latent attributes. IRT has been widely used in education and psychology. It can be used to develop or refine scales to measure latent traits. Furthermore, it offers a methodology for creating scales with desirable measurement properties. For example, if data fits an IRT model, a scale with invariant measurement properties has been developed. IRT also provides a way to examine the quality and appropriateness of items that can be used to measure what they are designed to measure. A benefit of IRT is that latent attributes of individuals and items can be inferred based on observed item responses. Furthermore, individuals and items can be placed and compared meaningfully on a latent scale. Due to this advantage, IRT has served as the foundation of computerised adaptive testing (CAT), which aims to accurately and efficiently assess individual's trait levels, such as in the Scholastic Aptitude Test (SAT) and

Graduate Record Examination (GRE).

As a statistical model, IRT has attracted attention in machine learning recently. Bergner et al. (2012) applied model-based collaborative filtering to estimate the parameters for IRT models, considering IRT as a type of collaborative filtering task, where the user-item interactions are factorised into user and item parameters. Bachrach et al. (2012) proposed a probabilistic graphical model that jointly models the difficulties of questions, the abilities of participants and the correct answers to questions in aptitude testing and crowdsourcing settings. While in MOOCs, Champaign et al. (2014) investigated the correlations between resource use and students' skill and relative skill improvement measured by IRT. Colvin et al. (2014) analysed pre-post test questions using IRT, to compare the learning in MOOCs and a blended on-campus course. Past work has tended to focus on using already-devised items to measure student ability under IRT models, while we are interested in automatically discovering content-based items that are characteristic of measurement in MOOCs (Milligan, 2015).

A variety of IRT models have been developed, and they differ from each other in terms of item characteristics, or item parameters, and the mathematical function (logistic or normal ogive curve) used for modelling the relationship between the characteristics of individuals and items, and the observed responses of individuals on items. The simplest IRT model is called the Rasch model or the One-Parameter Logistic Model (1PL), where the characteristic of an item is simply the item difficulty.

2.4.1 The Rasch Model

The Rasch model (Wright and Masters, 1982; Bond and Fox, 2001) for dichotomous data (correct/incorrect, agree/disagree responses) specifies the probability of a person's positive response (correct, agree) on an item as a logistic function of the difference between the person's ability and item difficulty, which can be formalised as

$$p_{ij} = P(X_{ij} = 1|\beta_i, \theta_j) = \frac{1}{1 + \exp\left(-\left(\theta_j - \beta_i\right)\right)} \ , \tag{2.2}$$

where θ_j denotes person j's ability, β_i denotes item i's difficulty, X_{ij} denotes the person j's response on item i, and p_{ij} denotes the probability of person j's positive response on item i. The probability can be illustrated by the Item Characteristic Curve (ICC) in Figure 2.2, commonly used in the field of IRT. It can be seen that the higher a person's ability relative to the difficulty of an item, the higher the probability of a positive response on that item. When a person's ability is equal to an item's difficulty on the latent scale, there is a 0.5 probability of a positive response on the item.

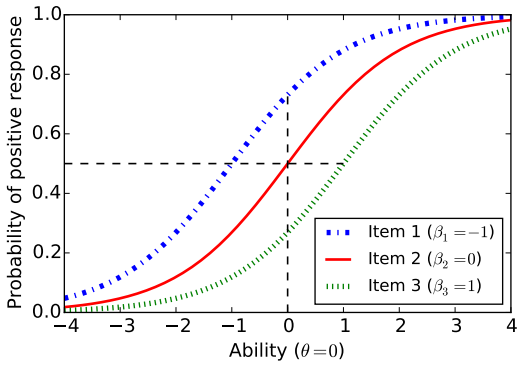


Figure 2.2: The Item Characteristic Curves for three items (item 1–the easiest, 3–the most difficult). A person with ability $\theta = 0$ has 0.5 probability to respond positively on item 2 with difficulty $\beta = 0$, and higher and lower probability on the easiest item 1 and the most difficult item 3 respectively.

The latent measurement scale is analogous to the ruler shown in Figure 2.3, where persons and items are placed and can be compared meaningfully. The Rasch model provides a way to construct the ruler using persons' responses on items. Persons and items are located along the scale according to their ability θ_j and difficulty β_i respectively.

The Rasch model, as a probabilistic model, can be viewed as a stochastic extension of the Guttman scale, with allowance for measurement error. For example, in Figure 2.3, person 1 and person 2 will have positive response on item 1 in a Guttman scale. While in a Rasch scale, there is a certain probability that the person 1 and person 2 will have positive response on item 1, and person 1 has higher probability compared to person 2. Such errors actually lead to a higher level of measurement scale — the interval scale, where we can tell how

much person 2 is better than person 1, not like the Guttman scale where we can only tell person 2 is better than person 1.

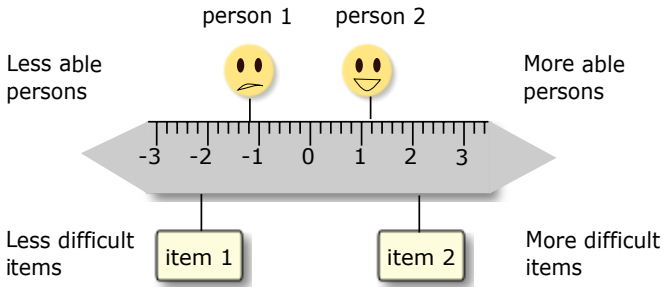


Figure 2.3: Representation of a latent ability scale. There is an increasing difficulty and ability from left to right.

Table 2.3 illustrates an example of items used for measuring basic mathematical ability with students' responses. Unlike the response pattern for a Guttman scale in Figure 2.2, a person might get a difficult item correct while getting an easy item incorrect in the Rasch. Each item difficulty and person ability is estimated on a logit scale. For example, person 1 responds positively 20% and negatively 80% on the items, and the person's initial ability is approximately -1.39 by taking the natural logarithm of the odds ratio for positive response $\frac{0.2}{0.8}$.

Table 2.3: An example of items for measuring basic mathematical ability, students' responses, initial item difficulty estimates and student ability estimates.

	Item 1 **(Count)**	**Item 2** **(+)**	**Item 3** **(−)**	**Item 4** **(×)**	**Item 5** **(÷)**	**Proportion correct** p_{θ_j}	**Ability** θ_j^0 $\log\left(\frac{p_{\theta_j}}{1-p_{\theta_j}}\right)$
Person 1	1	0	0	0	0	0.20	-1.39
Person 2	1	1	0	0	0	0.60	0.41
Person 3	0	1	1	0	0	0.60	0.41
Person 4	1	0	1	1	0	0.67	0.71
Person 5	1	1	1	0	1	0.80	1.39
Proportion correct p_{β_i}	0.80	0.33	0.33	0.20	0.20		
Difficulty β_i^0 $\log\left(\frac{1-p_{\beta_i}}{p_{\beta_i}}\right)$	-1.39	0.71	0.71	1.39	1.39		

Rasch Estimation

Given an observed response matrix $\mathbf{x}=[x_{ij}]$ (*e.g.*, Table 2.3), the goal is to estimate the person and item parameters θ_j and β_i. The most common estimation methods are based on maximum likelihood estimation: jointly maximum likelihood (JML) estimation, conditional maximum likelihood (CML) estimation and marginal maximum likelihood (MML) estimation (Baker and Kim, 2004). In this chapter, we focus on JML.

Under the assumption that a sample of n persons are drawn independently at random from a population of persons possessing the latent attribute, and the assumption of local independence that a person's responses to different items are statistically independent, the probability of an observed data matrix $\mathbf{x} = [x_{ij}]$ with k items and n persons is the product of the probabilities of the individual responses, and can be given by the likelihood function below:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{x}) &= \prod_{i=1}^{k}\prod_{j=1}^{n} P(X_{ij}=1|\beta_i,\theta_j)^{x_{ij}} \left(1 - P\left(X_{ij}=1|\beta_i,\theta_j\right)\right)^{(1-x_{ij})} \\ &= \prod_{i=1}^{k}\prod_{j=1}^{n} \frac{\exp\left(x_{ij}\left(\theta_j-\beta_i\right)\right)}{1+\exp(\theta_j-\beta_i)} \quad .\end{aligned} \tag{2.3}$$

Taking the logarithm of the likelihood function, we have

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}(\theta_j-\beta_i) - \sum_{i=1}^{k}\sum_{j=1}^{n} \log(1+\exp\left(\theta_j-\beta_i\right)) \quad . \tag{2.4}$$

The parameters of the Rasch model can be estimated by joint maximum likelihood using the iterative Newton-Raphson method, which yields the following iterative solution for βi and θj,

$$\beta_i^{t+1} = \beta_i^t - \frac{\sum_{j=1}^{n}(p_{ij}-x_{ij})}{-\sum_{j=1}^{n} p_{ij}(1-p_{ij})} \quad \text{for} \quad t \geq 0 \ , \tag{2.5}$$

$$\theta_j^{t+1} = \theta_j^t - \frac{\sum_{i=1}^{k}(x_{ij}-p_{ij})}{-\sum_{i=1}^{k} p_{ij}(1-p_{ij})} \quad \text{for} \quad t \geq 0 \ . \tag{2.6}$$

The convergence to a local optimum (with suitable step sizes) is guaranteed. The initial estimate of θ_j, θ_j^0 can be obtained by firstly calculating the proportion of items that a person j responded correctly p_{θ_j}, and then taking the natural logarithm of the odds of person j's correct response as shown in Table 2.3, which can be formalised as follows:

$$\theta_j^0 = \log\left(\frac{p_{\theta_j}}{1-p_{\theta_j}}\right), \quad p_{\theta_j} = \frac{r_j}{k}, \quad r_j = \sum_{i=1}^{k} x_{ij} \ , \tag{2.7}$$

where r_j denotes the number of items that person j responsed positively. Similarly, the initial estimate of β_i, β_i^0 can be obtained by

$$\beta_i^0 = \log\left(\frac{1-p_{\beta_i}}{p_{\beta_i}}\right), \quad p_{\beta_i} = \frac{s_i}{n}, \quad s_i = \sum_{j=1}^{n} x_{ij} \ , \tag{2.8}$$

where s_i denotes the number of persons who responded correctly on item i, and p_{β_i} denotes the proportion of persons who responded correctly on item i.

For those items that no one achieves a correct response ($s_i = 0$) on, or no one is incorrect on ($s_i = n$), some implementations of the Rasch model delete all of them, while other models handle the situation as follows (Baker and Kim, 2004):

$$s_i = \begin{cases} \epsilon, & \text{if } s_i = 0 \\ n - \epsilon, & \text{if } s_i = n \end{cases} ,$$

where ϵ is a small number, usually 0.5. Similar processing can be done for θ_j,

$$r_j = \begin{cases} \epsilon, & \text{if } s_i = 0 \\ k - \epsilon, & \text{if } s_i = k \end{cases} .$$

These pseudo counts are similar to frequentist Laplace corrections, or uniform Bayesian priors.

Evaluating Model Fit

The items can be said to measure the latent attribute on an interval scale when there is a close fit between data and model. The model-data fit is typically examined using Infit and Outfit statistics, which are two types of mean square error statistics, indicating information about the error in the estimates for the individual item and person.

Outfit and Infit test statistics are defined for each item and person, to test the fit of items and persons to the Rasch model, by summarising the Rasch residuals in different ways. The Rasch residuals are the differences between the observed responses and the expected responses according to the Rasch model. Formally, the expected response of person j on item i under the Rasch model $E(x_{ij})$ (abbreviated to E_{ij}) is $E(X_{ij}) = p_{ij}$. The residual between the observation x_{ij} and the expected response E_{ij} is $R_{ij} = x_{ij} - E_{ij}$. Standardised residuals are often used to assess the fit of a single person-item response

$$Z_{ij} = \frac{X_{ij} - E_{ij}}{\sqrt{Var(X_{ij} - E_{ij})}} = \frac{R_{ij}}{\sqrt{Var(X_{ij})}} \ , \tag{2.9}$$

where $Var(X_{ij}) = p_{ij}(1 - p_{ij})$ denotes the variance of X_{ij} (abbreviated to Var_{ij}).

The outfit of item i summarises the squared standardised residuals over persons, divided by the number of persons n, and is given by

$$\text{Outfit}_i = \frac{1}{n}\sum_{1}^{n} Z_{ij}^2 = \frac{1}{n}\sum_{1}^{n} \frac{R_{ij}^2}{Var_{ij}} \ . \tag{2.10}$$

Typical treatments assume standardized residuals Z_{ij} approximately following a unit normal distribution. Their sum of squares therefore approximately follows a χ^2 distribution. Dividing this sum by its degrees of freedom yields a mean-square value, with an expectation of 1.0 and taking values in the range of 0 to infinity.

Outfit is sensitive to unexpected responses to items that are relatively too easy or too hard for a person and vice-versa, *e.g.*, lucky guesses (*i.e.*, a person's responses 111001) and careless mistakes (*i.e.*, a person's responses 011100)

(Linacre, 2002). Since outfit is sensitive to the very unexpected observations (outliers), infit was devised to be more sensitive to the overall pattern of responses (Linacre, 2006). Infit is an information-weighted form of outfit: it weights the observations by their statistical information (model variance) which is larger for targeted observations, and smaller for extreme observations (Bond and Fox, 2001). In this thesis, we focus on infit. Formally, the infit of item i is given by

$$\text{Infit}_i = \frac{\sum_{j=1}^{n} Var_{ij} Z_{ij}^2}{\sum_{j=1}^{n} Var_{ij}} = \frac{\sum_{j=1}^{n} R_{ij}^2}{\sum_{j=1}^{n} Var_{ij}} \ . \tag{2.11}$$

Both outfit and infit have an expected value of 1.0. Values larger than 1.0 indicate underfit to the Rasch model, *i.e.*, the data are less predictable than the model expects, while values less than 1.0 indicate overfit of the data to the model, *i.e.*,, the observations are too predictable (Wright et al., 1994). Conventionally, the acceptable range is usually [0.7,1.3] or [0.8,1.2] depending on application.

2.5 Measurement Research in MOOCs and Machine Learning

As a statistical model, IRT has attracted attention in machine learning recently. Bergner et al. (2012) applied model-based collaborative filtering to estimate the parameters for IRT models, considering IRT as a type of collaborative filtering task, where the user-item interactions are factorised into user and item parameters. Bachrach et al. (2012) proposed a new probabilistic graphical model that jointly models the difficulties of questions, the abilities of participants and the correct answers to questions in aptitude testing and crowdsourcing settings.

While in MOOCs, Champaign et al. (2014) investigated the correlations between resource use and students' skill and relative skill improvement measured by IRT. Colvin et al. (2014) analysed the pre-post test questions using IRT, to compare the learning in MOOCs and a blended on-campus course. Past works focus on using already devised items to measure student ability under IRT models, while we are interested in automatically devising items based on forum

contents, which is characteristic of measurement in MOOCs (Milligan, 2015).

In traditional measurement, a latent attribute to be measured (*e.g.*, mathematical ability, reading skill) is first defined, and then a set of items (*e.g.*, questions) are devised and calibrated (e.g., deleted or changed) manually until persons' responses on such items fit a measurement model as evidence of reliability. In this thesis, we take a different perspective on measurement, as a task within an exploratory study. We automatically devise items (topics) from MOOC discussion forum content such that students' participation fits a measurement model for measuring student academic ability. In order to make sure that the discovered topic can be used for measuring student academic ability instead of other abilities or skills (*i.e.*, reading skill), we use students' grades as an indicator of their academic ability, and add constraints on their response patterns on items. In the end, we automatically devise items such that students' responses on these items satisfy the above constraints as evidence of measuring their academic ability.

Chapter 3

Identifying At-Risk Students in MOOCs

In this chapter, we explore the accurate and early identification of students who are at risk of not completing courses, and envision student interventions that present meaningful probabilities of success/failure, enacted only for marginal students. This intervention requires smoothed probabilities across weeks to make it effective. Based on regularised logistic regression, we propose two transfer learning algorithms to balance accuracy with smoothness. This chapter is based on the following publication: Jiazhen He, James Bailey, Benjamin I.P. Rubinstein, and Rui Zhang. "Identifying At-Risk Students in Massive Open Online Courses". In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1749–1755. AAAI Press, 2015.

3.1 Introduction

Identifying students who are at risk of not completing course is an important step for improving completion rates. Early prediction can help instructors design interventions to encourage course completion before a student falls too far behind. We focus on Coursera MOOCs, which often last for several weeks with students engaging in activities such as watching/downloading lectures, attempting assignments/quizzes, and posting to/viewing discussion forums.

To obtain early predictions, we build models weekly and leverage multiple offerings of a course to obtain ground truth to supervise the training of our models. Exploration of predictive analysis on MOOCs across multiple offerings has been limited thus far, but is nonetheless important, since data distributions across offerings is likely non-stationary: *e.g.*, different cohorts of students enrol in offerings, and course materials (lectures and assignments) are refined over time. It is not clear *a priori* whether a model trained on previous offerings will serve a new offering well.

A key aspect of our approach is a plan for interventions that involve presenting at-risk students with meaningful probabilities of success/failure. We hypothesise that such carefully crafted interventions could help students become aware of their progress and potentially persist. However a necessary condition for such an approach to be effective, is to have probabilities that are well calibrated. By focusing on intervening with only those students near the pass/fail borderline, we aim for students who could be motivated by being 'nearly there' in succeeding in the class. Our intervention plan expressly avoids displaying failure probabilities for high-risk students, for fear of discouraging them from further participation in the course. Therefore calibration is not necessary across the entire unit interval, only near 0.5.

By examining individual students' failure-probability trajectories, we observe huge fluctuations across weeks, which is undesirable for a number of reasons, such as confusing students or undermining credibility of the intervention system. Therefore, we impose a requirement of smoothed probabilities across consecutive weeks. Towards this end, we propose two transfer learning algorithms—Sequentially Smoothed Logistic Regression (LR-SEQ) and Simultaneously Smoothed Logistic Regression (LR-SIM)—to balance accuracy with smoothness. These algorithms add a regularisation term, which takes the probabilities in consecutive weeks into account, so as to minimise their difference. While LR-SEQ uses knowledge from the previous week to smooth the current week in a sequential fashion, LR-SIM learns across weeks simultaneously.

Contributions. The main contributions of this chapter are:

- The first exploration of early and accurate prediction of students at risk

of not completing a MOOC, with evaluation on multiple offerings, under potentially non-stationary data;

- An intervention that presents marginal students with meaningful success/failure probabilities: a novel approach to completion rates;

- Two transfer learning logistic regression algorithms which would be practical for deployment in MOOCs, for balancing accuracy & inter-week smoothness. Training converges quickly to a global optimum in both cases; and

- Experiments on two offerings of a Coursera MOOC that show the effectiveness of our algorithms in terms of accuracy, inter-week smoothness and calibration.

3.2 Problem Statement

We explore the accurate and early prediction of students who are at risk of failing, which we cast as a supervised binary classification task where possible class labels are whether or not a student will fail a course.

Predicted probabilities can serve a dual purpose, both for the identification of at-risk students and within subsequent intervention. We hypothesise that carefully employing the predicted probabilities as part of an intervention message could incentisise students to invest further in the course. Specifically, we propose to intervene with those who are on the pass/fail borderline rather than high-risk students. For example, given a 0.45 predicted probability, a hypothetical intervention message might resemble the following.

> Great work on your efforts so far—you're nearly there! In fact our statistical models suggest your profile matches students with a 55% chance of passing. This is based mainly on your lecture downloads this week. We'd like to encourage you to watch lecture 4 and post to the board. Doing just these 2 activities have greatly improved outcomes for students like you!

By targeting only those students near the pass/fail border, we are focusing on the part of the cohort that with an incremental investment could most personally benefit and increase the course pass rate.

Our application motivates 4 requirements of the learner.

- **Early & accurate predictions** enable timely interventions for at-risk students, with minimal unfounded and missing interventions;
- **Well-calibrated probabilities** allow proper targeting of interventions to those students who are truly near the classifier's decision boundary and to supply meaningful interventions: *e.g.*, approximately 60% of students with a risk prediction of 0.6 should eventually fail the course;
- **Smoothed probabilities** across consecutive weeks mitigate large fluctuations from slight changes in activities. Such fluctuations (*cf.* Figure 3.1) may undermine the credibility of intervention messages—we opt for consistent feedback. Moreover smoothing admits a principled approach to learning from the entire course when distributions change and even feature spaces change (*i.e.*, a form of regularisation through transfer learning); and
- **Interpretable models** suggest additions to intervention messages such as explanations for the current prediction and possible areas for improvement. Moreover such models can be useful in informing instructors on the profiles of successful vs. struggling students.

3.3 Algorithms

In initial experiments we explored a variety of supervised binary classifiers for predicting failure weekly: regularised logistic regression, SVM (LibSVM), random forest, decision tree (J48), naive Bayes, and BayesNet (in weka with default parameters used). Table 3.1 shows the results, indicating that regularised logistic regression performs best in terms of Area Under the ROC Curve (AUC), followed by BayesNet, naive Bayes, random forest, decision tree and SVM. Only

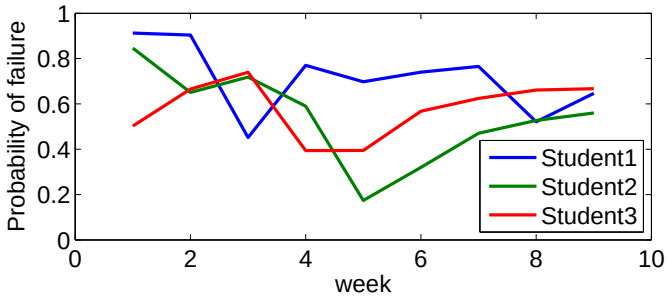


Figure 3.1: Failure-probability trajectories for three students across nine weeks produced by logistic regression with cross-validation performed weekly on *DisOpt* launched in 2014.

BayesNet is comparable to logistic regression, whilst SVM performs worst. In addition to the advantage of outperforming other classifiers, logistic regression: produces interpretable linear classifiers with weights indicating relative importance (under certain assumptions); is naturally well-calibrated (Niculescu-Mizil and Caruana, 2005b); and is a technique widely appreciated by researchers in the education community. Therefore in the sequel we focus our attention on approaches based on logistic regression.

Table 3.1: Comparison of different classifiers in term of AUC across 9 weeks on DisOpt.

	1	2	3	4	5	6	7	8	9
DecisionTree	0.716	0.759	0.847	0.873	0.921	0.909	0.948	0.958	1.000
RandomForest	0.689	0.825	0.872	0.910	0.938	0.953	0.972	0.983	1.000
NaiveBayes	0.764	0.858	0.909	0.938	0.956	0.968	0.979	0.984	0.987
BayesNet	0.765	0.871	0.919	0.947	0.962	0.975	0.986	0.993	1.000
LR	**0.771**	**0.875**	**0.922**	**0.950**	**0.967**	**0.978**	**0.986**	**0.993**	**0.999**
SVM	0.537	0.638	0.732	0.786	0.827	0.879	0.905	0.936	0.978

To address smoothness, we propose two adaptations to logistic regression. To aid their development, we first review basic regularised logistic regression. A glossary of symbols used in this chapter is given in Table 3.2.

Table 3.2: Glossary of symbols

Symbol	Description
n	The number of weeks
n_i	The number of students by week i
$n_{i,i-1}$	The number of extant students by both week i and week i-1
$\boldsymbol{x}_i$	The set of students by week i
$\boldsymbol{x}_{ij}$	The jth student by week i
d_i	The number of features for student $\boldsymbol{x}_{ij}$
$\boldsymbol{x}_i^{(i-1,i)}$	The set of students in week i also existing in week i-1
$\boldsymbol{x}'_{ij}$	The jth student with extended feature space by week i
$\boldsymbol{x}_i'^{(i-1,i)}$	The set of students with extended feature space by week i also existing in week i-1
$\boldsymbol{w}_i$	The weight vector for week i
$\boldsymbol{w}$	The weight vector for all weeks
$\boldsymbol{y}_i$	The set of labels for students by week i
$\boldsymbol{y}_{ij}$	The label of jth student by week i
λ_1	Regularisation parameter for overfitting
λ_2	Regularisation parameter for smoothness

3.3.1 Logistic Regression (LR)

Let n be the number of weeks that a course lasts for. We have n_i students by the end of week i ($1 \leq i \leq n$). $\boldsymbol{x}_i = \{\boldsymbol{x}_{i1}, \boldsymbol{x}_{i2}, \cdots, \boldsymbol{x}_{in_i}\}$ is the set of students in week i. Each student $\boldsymbol{x}_{ij}$ is described by d_i features. Note that the number of students by the end of each week i can be different, since students can enter a course at any time while it is running.

Logistic regression predicts label y (fail for y=1 and pass for y=-1) for input vector $\boldsymbol{x}_{ij}$ (a student) according to,

$$p(y|\boldsymbol{x}_{ij}, \boldsymbol{w}_i) \quad = \quad \sigma(y\boldsymbol{w}_i^T\boldsymbol{x}_{ij}) = \frac{1}{1+\exp(-y\boldsymbol{w}_i^T\boldsymbol{x}_{ij})} \tag{3.1}$$

where $\boldsymbol{w}_i = [w_{i1}, w_{i2}, \cdots, w_{id_i}]^T$ is the weight vector to be learned.

From a data set by week i, given by $(\boldsymbol{x}_i, \boldsymbol{y}_i) = [(\boldsymbol{x}_{i1}, y_{i1}), (\boldsymbol{x}_{i2}, y_{i2}), \cdots, (\boldsymbol{x}_{in_i}, y_{in_i})]$,

we wish to find $\boldsymbol{w}_i$ by L_2-regularised maximum likelihood estimation: minimising with regularisation parameter $\lambda_1 > 0$,

$$\mathcal{L}(\boldsymbol{w}_i) = \sum_{j=1}^{n_i} \log(1 + \exp(-y_{ij}\boldsymbol{w}_i^T x_{ij})) + \frac{\lambda_1}{2} \|\boldsymbol{w}_i\|^2 \tag{3.2}$$

This convex problem can be solved by Newton-Raphson which produces the update equations known as iteratively-reweighted least squares. The result is n logistic regression models learned separately by the end of each week.

3.3.2 Sequentially Smoothed LR (LR-SEQ)

In order to smooth probabilities across weeks, we propose a transfer learning algorithm, *Sequentially Smoothed Logistic Regression (LR-SEQ)*. Transfer learning leverages the knowledge learned in related tasks to better learn a new task. In our setting, the previous week's knowledge is used to help learn smoothed probabilities for the current week.

A natural approach is to follow existing transfer learning approaches to linear classifiers (Ando and Zhang, 2005): add a regularisation term minimising the difference between $\boldsymbol{w}_i$ and $\boldsymbol{w}_{i-1}$. However, the data distribution across weeks can be non-stationary as engagement varies and prescribed activities evolve. Moreover the number of features might change ($d_i \neq d_{i-1}$). Instead we seek to minimise the difference between predicted probabilities between two weeks directly. Unfortunately this leads to a non-convex objective. Therefore we minimise a surrogate: the difference between $\boldsymbol{w}_i \boldsymbol{x}_i^{(i-1,i)}$ and $\boldsymbol{w}_{i-1} \boldsymbol{x}_{i-1}^{(i-1,i)}$, where $\boldsymbol{x}_i^{(i-1,i)}$ denotes the set of students in week i that also exist in week $i-1$, and similarly $\boldsymbol{x}_{i-1}^{(i-1,i)}$ denotes the set of students in week $i-1$ that also exist in

week i. The objective function[1] for week i is

$$\begin{aligned}\mathcal{L}(\boldsymbol{w}_i) &= \sum_{j=1}^{n_i} \log(1+\exp(-y_j \boldsymbol{w}_i^T \boldsymbol{x}_{ij})) + \frac{\lambda_1}{2} \|\boldsymbol{w}_i\|^2 \\ &+ \lambda_2 \sum_{j=1}^{n_{i,i-1}} \left\| \boldsymbol{w}_i^T \boldsymbol{x}_{ij}^{(i,i-1)} - \boldsymbol{w}_{i-1}^T \boldsymbol{x}_{i-1j}^{(i,i-1)} \right\|^2 \end{aligned} \tag{3.3}$$

where parameter $\lambda_2 > 0$ controls smoothness and the level of transfer. This surrogate objective function is convex therefore efficiently solved by Newton-Raphson to a guaranteed global optimum. To recap: n weekly logistic regression models are learned sequentially such that week i's model cannot be built until model for week $i-1$ is obtained.

3.3.3 Simultaneously Smoothed LR (LR-SIM)

The drawback of LR-SEQ is that early inaccurate predictions cannot benefit from the knowledge learned in later weeks (where data is closer to the end of the course), in-turn undermining models learned later. To combat this effect, we propose *Simultaneously Smoothed Logistic Regression (LR-SIM)* that simultaneously learns models for all weeks. LR-SIM allows early and later prediction to be correlated and to influence each other, which we expect should yield improved prediction due to inter-task regularisation but also good smoothness.

We first extend the feature space for each student $\boldsymbol{x}_{ij}$ to a new space with n components. The student $\boldsymbol{x}'_{ij}$ with new feature space has $\sum_{i=1}^{n} d_i$ dimensions, with the ith component having d_i features corresponding to the features in the original feature space by the end of week i, and others zero. For example, for a student at the end of week 2, $\boldsymbol{x}_{2j}$, we extend to a new point $\boldsymbol{x}'_{2j}$, where the 2nd component with d_2 features are actually the same as $\boldsymbol{x}_{2j}$, and others being zero. Hence we encode the same information by the end of week 2 that contributes to the outcome. We must learn a single $\boldsymbol{w}$, which also has $\sum_{i=1}^{n} d_i$ dimensions corresponding to $\boldsymbol{x}'_{ij}$. But only the ith component—the ith model—contributes to the prediction by the end of week i, due to the zero values of other dimensions

[1] For $i \geq 2$; the objective for week 1 is identical to LR in Eq. (3.2).

of $\boldsymbol{x}_{ij}^{'}$.

$$\begin{array}{c} \\ \boldsymbol{x}_{1j}^{'} \\ \boldsymbol{x}_{2j}^{'} \\ \vdots \\ \boldsymbol{x}_{nj}^{'} \end{array} \begin{array}{c} \begin{array}{cccc} 1 & 2 & \cdots & n \end{array} \\ \begin{pmatrix} \boldsymbol{x}_{1j} & [0,\cdots,0] & \cdots & [0,\cdots,0] \\ [0,\cdots,0] & \boldsymbol{x}_{2j} & \cdots & [0,\cdots,0] \\ \vdots & \vdots & \ddots & \vdots \\ [0,\cdots,0] & [0,\cdots,0] & \cdots & \boldsymbol{x}_{nj} \end{pmatrix} \end{array}$$

Based on the extended $\boldsymbol{x}_{ij}^{'}$ and $\boldsymbol{w}$, we can minimise the difference of probabilities predicted for week i and week $i-1$ for i $(i \geq 2)$ together, via a simple expression, as shown in Eq. (3.4). Again the objective function is convex and efficiently minimised.

$$\begin{aligned} \mathcal{L}(\boldsymbol{w}) = & \sum_{i=1}^{n} \sum_{j=1}^{n_i} \log(1+\exp(-y_j \boldsymbol{w}^T \boldsymbol{x}_{ij}^{'}) + \frac{\lambda_1}{2} \|\boldsymbol{w}\|^2 \\ & + \lambda_2 \sum_{i=2}^{n} \sum_{j=1}^{n_{i,i-1}} \left\| \boldsymbol{w}^T \boldsymbol{x}_{ij}^{'(i,i-1)} - \boldsymbol{w}^T \boldsymbol{x}_{i-1j}^{'(i,i-1)} \right\|^2 \end{aligned} \tag{3.4}$$

Our algorithms can operate for tasks with differing feature spaces and feature dimensions. For example, one might use individual-level features for each lecture and assignment, which might be released weekly, to help understand and interpret student performance.

3.4 Experimental Results

We conduct experiments to evaluate the effectiveness of our algorithms on real MOOCs.

3.4.1 Dataset Preparation

Discrete Optimization Dataset

The first offering of *Discrete Optimization* (***DisOpt1***) launched in 2013 by The University of Melbourne lasted for nine weeks, with 51,306 students enrolled, of which 795 students received a certificate of completion for the course. This course has an open course curriculum with all the videos and assignments released at the beginning of the course, enabling students to study at their own pace. There are 57 video lectures and 7 assignments in total. Students can watch/download video lectures, and attempt assignments multiple times. Their final grade is assessed by the total score on 7 assignments.

The second offering of *Discrete Optimization* (***DisOpt2***) launched in 2014 also lasted for nine weeks, attracting 33,975 students to enroll, of which 322 students completed. There are 4 fewer video lectures compared to *DisOpt1*, with 43 video lectures. The number of assignments remain but some of the assignment contents differ to those of *DisOpt1*. The total score of all assignments differs between offerings. An overview of the two offerings is shown in Table 3.3.

Table 3.3: Overview on two offerings for *DisOpt*

	DisOpt1	*DisOpt2*
Duration	9 weeks	9 weeks
Number of students enrolled	51,306	33,975
Number of students completed	795	322
Number of video lectures	57	53
Number of assignments	7	7
Total score of all assignments	396	382

Cohorts

Among all the students enrolled, only a tiny fraction complete, which makes the data extremely imbalanced. Figure 3.2 shows the number of students in different course activities. In *DisOpt1*, among all the students enrolled, only around 41%, 13% and 2% of the students watch/download videos, do assignments and

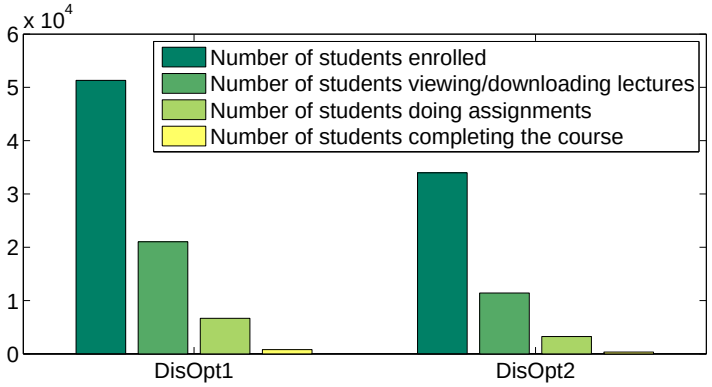


Figure 3.2: Student participation in the first and second offering of *Discrete Optimization MOOC*

complete the course respectively. The same thing happens in *DisOpt2* with low completion rate, and *DisOpt2* had fewer students enrolled. Students enroll for various reasons. For example, some treat MOOCs like traditional courses by taking lectures and assignments at a fixed pace, while others treat MOOCs as online references without doing any assignments (Anderson et al., 2014). In this chapter, we are interested in helping those who intend to pass. Therefore, we focus on students who are active in assignments/quizzes, which indicates an intention to pass. In particular, at the end of each week, we retain the students who did at least one assignment by that week.

Features Used

We extract features from student engagement with video lectures and assignments, and performance on assignments by the end of each week to predict their performance at the end of the course. The features are shown in Table 3.4. In order to easily apply the model trained on previous offerings to a new offering, we extract features present across offerings.

3.4.2 Performance Measure

To evaluate the effectiveness of our proposed algorithms, we train prediction models on *DisOpt1*, and test on *DisOpt2*. Due to the class imbalance where high

Table 3.4: Features for each week i for *DisOpt*

Features
Percentage of lectures viewed/downloaded by week i
Percentage of lectures viewed/downloaded in week i
Percentage of assignments done by week i
Percentage of assignments done in week i
Average attempts on each assignment done by week i
Average attempts on each assignment done in week i
Percentage of score on assignments done by week i, to total score on all assignments

proportion of students fail, we prefer area under the ROC curve (AUC), which is invariant to imbalance. To measure the smoothness for week i, we compute the difference of probabilities between week i and week i-1 for each active student (in terms of our rule for maintaining students) in week i and i-1, and obtain the averaged difference for all students, and standard deviation (stdev).

3.4.3 Smoothness and AUC

To evaluate the effectiveness of our proposed algorithms LR-SEQ and LR-SIM, we compare them with two baselines, LR and a simple method using moving averages, denoted LR-MOV. LR-MOV predicts as final probability for week i an average of LR's week i and i-1 probabilities, $(i \geq 2)$. The prediction for week 1 is the same as LR. We train models using the above four algorithms on $DisOpt1$, where $\lambda_1 = 10$ and $\lambda_2 = 1$, and apply them to $DisOpt2$. Figure 3.3 and Table 3.5 show the smoothness and AUC across weeks respectively.

As we can see from Figure 3.3, LR-SEQ and LR-SIM achieve better smoothness (average difference) and low standard deviation, especially in the first five weeks where early intervention is most critical. LR attains smooth probabilities in the last few weeks, but with high standard deviation, when intervention is less impactful. LR-MOV achieves the same smoothness as LR with reduced standard deviation, demonstrating the need for performing some kind of smoothing.

From Table 3.5, we can see that LR-SIM and LR-MOV are comparable to

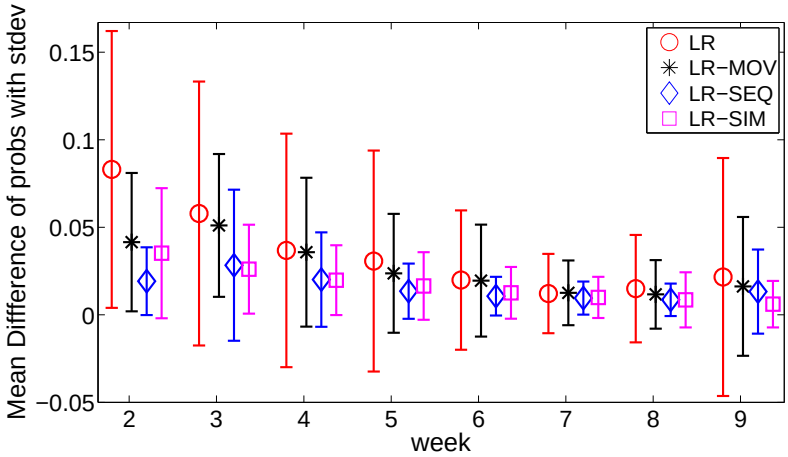


Figure 3.3: Comparison of LR, LR-MOV, LR-SEQ and LR-SIM on smoothness across weeks. Mean difference of probabilities across students plus/minus standard deviation. Closer to zero difference is better.

Table 3.5: Comparison of LR, LR-MOV, LR-SEQ and LR-SIM on AUC across weeks.

Week	LR	LR-MOV	LR-SEQ	LR-SIM
1	0.788	0.788	0.788	**0.800**
2	0.867	0.856	0.849	**0.872**
3	0.901	0.890	0.867	0.892
4	0.928	0.923	0.907	0.923
5	0.947	0.944	0.934	0.944
6	0.962	0.958	0.953	0.960
7	0.970	0.968	0.963	0.969
8	0.984	0.981	0.981	0.986
9	0.996	0.997	0.997	0.995

LR in terms of AUC, while LR-SEQ decreases slightly. (Note: LR-MOV cannot achieve better smoothness as shown in Figure 3.3.) LR-SIM does outperform LR in the first two weeks (in bold): one reason might be that the reduced model complexity due to transfer learning helps to mitigate overfitting; another rea-

son might be that later, more accurate predictions improve early predictions via transfer learning in *DisOpt1* and the data distributions over *DisOpt1* and *DisOpt2* do not significantly vary. On the other hand, LR-SEQ gets continually worse in the first three weeks: LR-SEQ only uses the previous week's knowledge to constrain the present week, but early predictions might be inaccurate, which undermine models learned later (*cf.* week 3, with the worst AUC). Later, LR-SEQ catches up with LR as data closer to the end of the course becomes available.

Overall, LR-SIM and LR-SEQ outperform LR consistently in terms of smoothness. And LR-SIM maintains or even improves on LR's AUC in early weeks, while LR-SEQ suffers slightly inferior AUC in the first few weeks, and is comparable to LR in the last few weeks. Notably, using the data collected by the end of early weeks we can achieve quite good AUC: about 0.87 by week 2 and 0.9 by week 3, *establishing the efficacy of early identification of at-risk students*. Furthermore, this demonstrates that a model trained on the first offering works well on the second offering.

3.4.4 Parameter Analysis

We compare the performance of LR-SIM, LR-SEQ and LR in terms of smoothness and AUC varying λ_1 and λ_2. Figure 3.4 shows results for week 2. We choose week 2 to emphasise early intervention. The curves from right to left show varying λ_2 from 10^{-4} to 10^4. The smoothness is computed between week 2 and week 1, and AUC is for week 2. It can be seen that LR achieves good AUC but poor smoothness. LR-SIM dominates LR-SEQ. As λ_2 increases, LR-SEQ and LR-SIM get smoother. But LR-SIM can achieve better AUC while LR-SEQ gets worse. Overall, LR-SIM clearly outperforms LR-SEQ and LR.

3.4.5 Calibration

Given an instance, it is not possible to know what the true underlying probability is, therefore some approximations are often used. A common way is to group instances based on the ranked predicted probability into deciles of

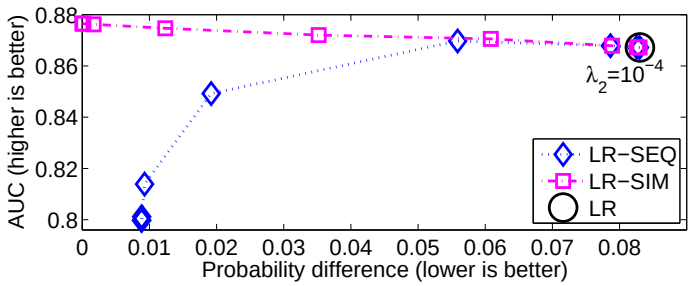


Figure 3.4: Smoothness versus AUC for LR, LR-SEQ and LR-SIM for week 2 when $\lambda_1 = 10$, and $\lambda_2 = 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}, 10^{3}, 10^{4}$.

risk with approximately equal number of instances in each group, and compare the predicted probability with observed probability within each group. A reliability diagram plotting the predicted probability with observed probability, is commonly used for calibration (Niculescu-Mizil and Caruana, 2005a; Zhong and Kwok, 2013).

Figure 3.5 shows the reliability diagram using LR-SIM for week 2. Our predicted probabilities agree closely with the observed probability in the gray region of marginal at-risk students for whom we wish to intervene.

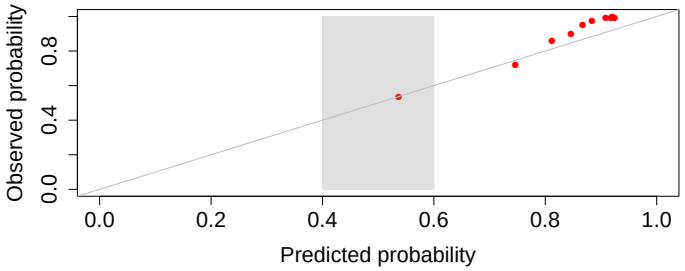


Figure 3.5: Reliability diagram for class *fail* using LR-SIM week 2. Grey area shows an intervention interval [0.4,0.6], which could be varied according to educational advice.

3.5 Conclusion

We have taken an initial step towards early and accurately identifying at-risk students, which can help instructors design interventions. We have compared different prediction models, with regularised logistic regression preferred due to its good performance, calibration and interpretability. Based on the predicted probabilities, we envision an intervention that presents students meaningful probabilities to help them realise their progress. We developed two novel transfer learning algorithms LR-SEQ and LR-SIM based on regularised logistic regression. Our experiments on Coursera MOOC data indicate that LR-SEQ and LR-SIM can produce smoothed probabilities while maintaining AUC, with LR-SIM outperforming LR-SEQ. LR-SIM has exceptional AUC in the first few weeks, which is promising for early prediction. Our experiments leveraging the two offerings of a Coursera MOOC demonstrate that the prediction models trained on a first offering work well on a second offering.

Chapter 4

Topic-Instrumented Measurement Based on the Guttman Scale

In this chapter, we explore the suitability of using automatically discovered topics from MOOC discussion forum content for modelling students' academic ability. If students' participation across the discovered topics fit a measurement model (in this chapter we adopt the Guttman scale) for measuring statistical effectiveness, and the topics are interpretable to subject-matter experts for measuring qualitative effectiveness, then the discovered topics can be regarded as useful items for measurement. In order to discover topics such that students' participation across them conforms to the Guttman scale, we introduce a novel regularisation into non-negative matrix factorisation-based topic modelling that incorporates the Guttman scale constraint. The resulting Guttman scaled topics could be used for student assessment and curriculum refinement. This chapter is based on the following publication: Jiazhen He, Benjamin I.P. Rubinstein, James Bailey, Rui Zhang, Sandra Milligan, and Jeffrey Chan. "MOOCs Meet Measurement Theory: A topic-Modelling Approach". In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1195–1201. AAAI Press, 2016.

4.1 Introduction

The lack of instructor-student interactivity has made assessment tools serve not only for grading, but also as an alternative and main way to provide feedback to students (Chauhan, 2014). Traditionally, quizzes or assignments are used as assessment tools to evaluate student proficiency. However, MOOCs with the rich student engagement data generated, provide opportunities and new perspectives of gaining insights into student learning and provide feedback. Recent research on educational measurement propose that a distinctive and complex skill is required to promote learning, and this skill can be evidenced by students' engagement behaviour patterns in MOOCs, not just visible assessment tools such as quizzes or assignments by themselves. Measurement theory has been used to build a reliable measure of this learning skill from student engagement data (Milligan, 2015). On the other hand, among the variety of engagement activities in MOOCs, MOOC discussion forums, as a platform and primary means of interaction, providing help and seeking help, produce rich content data. Understanding the discussions can be helpful for understanding student learning and providing feedback.

Inspired by the recent research on educational measurement, together with the importance of MOOC discussion forums for understanding student learning and providing feedback, we explore the problem that using students' participation across automatic discovered topics as evidence of measuring student ability. In MOOCs, as in the traditional classroom, we may hypothesise that students possess a latent ability in the subject at hand. For example, in a MOOC on macroeconomics, students are expected to develop knowledge in introductory macroeconomics via videos, quizzes and forums. Students' latent abilities can be defined, validated and measured using indicators drawn from student responses to activities like interaction with videos, quiz results and forum participation. In this chapter,we focus on using the content of forum discussion in MOOCs for measurement, which is too time-consuming to analyse manually but that can provide a predictive indicator of achievement (Beaudoin, 2002).

In order to validate such a hypothesis, a measurement model can be used as evidence as to whether automatically discovered topics are appropriate for

measuring student ability. In this chapter, we use the Guttman scale. In particular, we automatically generate items (topics) from unstructured forum data using topic modelling. The topics are generated in such a way that the students' dichotomous response on the topics (posting on a topic or not) are enforced to conform to a Guttman scale. This establishes the statistical effectiveness of using topics as an instrument to measure student ability. In addition, the topics are required to be interpretable to subject-matter experts who could be teaching such MOOCs. For example, for a MOOC on discrete optimisation, our goal is to automatically discover topics such as *How to use platform/python*—the easiest which most students contribute to—and *How to design and tune simulated annealing and local search*—a more difficult topic which only a few students might post on. Such well-scaled topics could be useful for student assessment, and the inferred topic difficulty levels could be used for curriculum refinement.

The challenge is to formalise the Guttman scale educational constraint and incorporate it into topic models. We opt to focus on non-negative matrix factorisation (NMF) approaches to topic modelling, as these admit natural integration of the Guttman scale educational constraint.

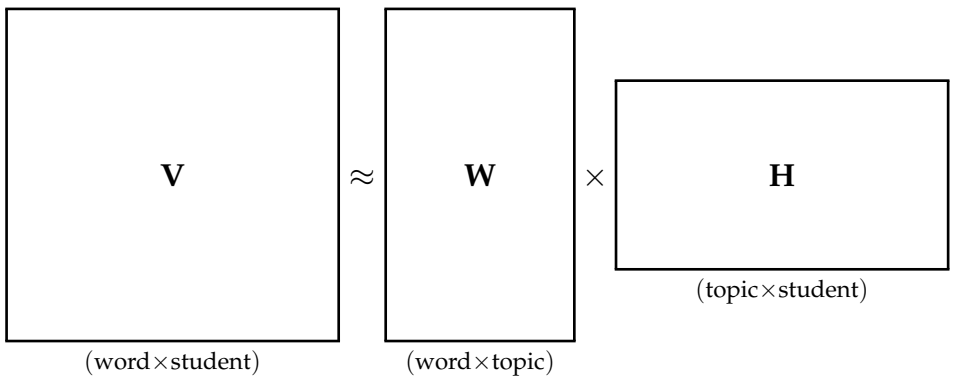
Contributions. The main contributions of this chapter are:

- A first study of how a machine learning technique, NMF-based topic modelling, can be used for the education research topic of psychometric testing;
- A novel regularisation of NMF that incorporates the educational constraint that inferred topics form a Guttman scale; and accompanying training algorithm;
- Quantitative experiments on three Coursera MOOCs covering a broad swath of disciplines, establishing statistical effectiveness of our algorithm; and
- A carefully designed qualitative survey of experts in two MOOC subjects, which supports the interpretability of our results and suggests their applicability in education.

4.2 Problem Statement

We explore the automatic discovery of forum discussion topics for measurement in MOOCs. Our central tenet is that topics can be regarded as useful items for measuring a latent skill, if student responses to these items conform to a Guttman scale, and if the topics are semantically-meaningful to domain experts. As Guttman scale item responses are typically dichotomous, we consider item responses to be whether a student posts on the topic or not. Our goal is to generate a set of meaningful topics that yield a student-topic matrix conforming to the properties of a *Guttman scale*, *e.g.*, a near-triangular matrix (see Table 2.2). This process can be cast as optimisation.

We choose NMF as the basic approach to discover forum topics due to the interpretability of topics produced, and the extensibility of its optimisation program. Using NMF, a word-student matrix $\mathbf{V}$ can be factorised into two non-negative matrices: word-topic matrix $\mathbf{W}$ and topic-student matrix $\mathbf{H}$ as follow:



Our application requires that the topic-student matrix $\mathbf{H}$ be **a) Binary** ensuring the response of a student to a topic is dichotomous; and **b) Guttman-scaled** ensuring the student responses to topics conform to a Guttman scale. NMF provides an elegant framework for incorporating these educational constraints via adding novel regularisation, as detailed in the next section. A glossary of important symbols used in this chapter is given in Table 4.1.

Table 4.1: Glossary of symbols

Symbol	Description
m	the number of words
n	the number of students
k	the number of topics
$\mathbf{V} = (v_{ij})_{m \times n}$	word-student matrix
$\mathbf{W} = (w_{ij})_{m \times k}$	word-topic matrix
$\mathbf{H} = (h_{ij})_{k \times n}$	topic-student matrix
$\mathbf{H}_{ideal} = ((h_{ideal})_{ij})_{k \times n}$	exemplar topic-student matrix with ideal Guttman scale
$\lambda_0, \lambda_1, \lambda_2$	regularisation coefficients

4.3 NMF for Guttman scale (NMF-Guttman)

4.3.1 Primal Program

We introduce the following regularisation terms on $\mathbf{W}$ to prevent overfitting, and on $\mathbf{H}$ to encourage a binary solution and Guttman scaling:

- $\|\mathbf{W}\|_F^2$ to prevent overfitting;
- $\|\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ to encourage a Guttman-scaled $\mathbf{H}$, where $\mathbf{H}_{ideal}$ is a constant matrix with ideal Guttman scale;
- $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ to encourage a binary solution $\mathbf{H}$, where operator $\circ$ denotes the Hadamard product.

Binary matrix factorisation (BMF) is a variation of NMF, where the input matrix and the two factorised matrices are all binary. Inspired by the approach of Zhang et al. (2007) and Zhang et al. (2010), we add regularisation term $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$. Noting this term equals $\|\mathbf{H} \circ (\mathbf{H} - \mathbf{1})\|_F^2$, it is clearly minimised by binary $\mathbf{H}$.

These terms together yield the objective function

$$f(\mathbf{W}, \mathbf{H}) = \|\mathbf{V} - \mathbf{W}\mathbf{H}\|_F^2 + \lambda_0 \|\mathbf{W}\|_F^2 + \lambda_1 \|\mathbf{H} - \mathbf{H}_{ideal}\|_F^2 + \lambda_2 \|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2 \ , \tag{4.1}$$

where $\lambda_0, \lambda_1, \lambda_2 > 0$ are regularisation parameters; with primal program

$$\min_{\mathbf{W},\mathbf{H}} f(\mathbf{W},\mathbf{H}) \quad \text{s.t.} \quad \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0} \ . \tag{4.2}$$

4.3.2 Algorithm

A local optimum of program (4.2) is achieved via iteration

$$w_{ij} \leftarrow w_{ij} \frac{(\mathbf{V}\mathbf{H}^T)_{ij}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T + \lambda_0 \mathbf{W})_{ij}} \tag{4.3}$$

$$h_{ij} \leftarrow h_{ij} \frac{(\mathbf{W}^T\mathbf{V})_{ij} + 4\lambda_2 h_{ij}^3 + 3\lambda_2 h_{ij}^2 + \lambda_1 (h_{ideal})_{ij}}{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + 6\lambda_2 h_{ij}^3 + (\lambda_1 + \lambda_2) h_{ij}} \tag{4.4}$$

These rules for the constrained program can be derived via the Karush-Kuhn-Tucker conditions necessary for local optimality. First we construct the unconstrained Lagrangian

$$\mathcal{L}(\mathbf{W},\mathbf{H},\boldsymbol{\alpha},\boldsymbol{\beta}) = f(\mathbf{W},\mathbf{H}) + \mathrm{tr}(\boldsymbol{\alpha}\mathbf{W}) + \mathrm{tr}(\boldsymbol{\beta}\mathbf{H}) \ ,$$

where $\alpha_{ij}, \beta_{ij} \leq 0$ are the Lagrangian dual variables for inequality constraints $w_{ij} \geq 0$ and $h_{ij} \geq 0$ respectively, and $\boldsymbol{\alpha} = [\alpha_{ij}]$, $\boldsymbol{\beta} = [\beta_{ij}]$ denote their corresponding matrices.

The KKT condition of stationarity requires that the derivative of $\mathcal{L}$ with respect to $\mathbf{W}, \mathbf{H}$ vanishes:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}} &= 2\left(\mathbf{W}^\star \mathbf{H}^\star \mathbf{H}^{\star T} - \mathbf{V}\mathbf{H}^{\star T} + \lambda_0 \mathbf{W}^\star\right) + \boldsymbol{\alpha}^\star = \mathbf{0} \ , \\
\frac{\partial \mathcal{L}}{\partial \mathbf{H}} &= 2\Big(\mathbf{W}^{\star T}\mathbf{W}^\star \mathbf{H}^\star - \mathbf{W}^{\star T}\mathbf{V} + (\lambda_1 + \lambda_2)\mathbf{H}^\star - \lambda_1 \mathbf{H}_{ideal}\Big) \\
&\quad + 4\lambda_2 \mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - 6\lambda_2 \mathbf{H}^\star \circ \mathbf{H}^\star + \boldsymbol{\beta}^\star = \mathbf{0} \ .
\end{aligned}$$

Complementary slackness $\alpha_{ij}^{\star} w_{ij}^{\star} = \beta_{ij}^{\star} h_{ij}^{\star} = 0$, implies:

$$\begin{aligned} 0 &= \left(\mathbf{V}\mathbf{H}^{\star T} - \mathbf{W}^{\star}\mathbf{H}^{\star}\mathbf{H}^{\star T} - \lambda_0 \mathbf{W}^{\star}\right)_{ij} w_{ij}^{\star} \ , \\ 0 &= \Big(\mathbf{W}^{\star T}\mathbf{V} + 3\lambda_2 \mathbf{H}^{\star} \circ \mathbf{H}^{\star} + \lambda_1 \mathbf{H}_{ideal} - \mathbf{W}^{\star T}\mathbf{W}^{\star}\mathbf{H}^{\star} - 2\lambda_2 \mathbf{H}^{\star} \circ \mathbf{H}^{\star} \circ \mathbf{H}^{\star} \\ &\quad -(\lambda_1 + \lambda_2)\mathbf{H}^{\star} + 4\lambda_2 \mathbf{H}^{\star} \circ \mathbf{H}^{\star} \circ \mathbf{H}^{\star} - 4\lambda_2 \mathbf{H}^{\star} \circ \mathbf{H}^{\star} \circ \mathbf{H}^{\star}\Big)_{ij} h_{ij}^{\star} \ . \end{aligned}$$

These two equations lead to the updating rules (4.3), (4.4). Our next result proves that these rules improve the objective value.

Theorem 4.1. *The objective function* $f(\mathbf{W}, \mathbf{H})$ *of program* (4.2) *is non-increasing under update rules* (4.3) *and* (4.4).

Proof. We follow the similar procedure described in (Lee and Seung, 2001), where an auxiliary function similar to that used in the Expectation-Maximization (EM) algorithm is used for proof.

Definition 4.2. *(Lee and Seung, 2001)* $G(h, h')$ *is an auxiliary function for* $F(h)$ *if the conditions*

$$G(h, h') \geq F(h), \qquad G(h, h) = F(h)$$

are satisfied.

Lemma 1. *(Lee and Seung, 2001) If G is an auxiliary function, then F is non-increasing under the update*

$$h_{t+1} = \underset{h}{\operatorname{argmin}}\, G(h, h^t) \tag{4.5}$$

Proof: $F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t)$

For any element h_{ij} in $\mathbf{H}$, let $F_{h_{ij}}$ denote the part of $f(\mathbf{W}, \mathbf{H})$ in Eq. (5.1) in the chapter relevant to h_{ij}. Since the update is essentially element-wise, it is sufficient to show that each $F_{h_{ij}}$ is non-increasing under the update rule of Eq. (4.4) in the chapter. To prove it, we define the auxiliary function regarding h_{ij} as follows.

Lemma 2. *Function*

$$G(h_{ij}, h_{ij}^t) = F_{h_{ij}}(h_{ij}^t) + F'_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t) + \varphi_{ij}(h_{ij} - h_{ij}^t)^2 \tag{4.6}$$

where

$$\varphi_{ij} = \frac{(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + \lambda_1 h_{ij}^t + 6\lambda_2 (h_{ij}^t)^3 + \lambda_2 h_{ij}^t}{h_{ij}^t}$$

is an auxiliary function for $F_{h_{ij}}$.

Proof: It is obvious that $G(h_{ij}, h_{ij}) = F_{h_{ij}}$. So we only need to prove that $G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$. Considering the Taylor series expansion of $F_{h_{ij}}$,

$$F_{h_{ij}} = F_{h_{ij}}(h_{ij}^t) + F'_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t) + \frac{1}{2}F''_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t)^2$$

$G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$ is equivalent to $\varphi_{ij} \geq \frac{1}{2}F''_{h_{ij}}(h_{ij}^t)$, where

$$F''_{h_{ij}}(h_{ij}^t) = 2(\mathbf{W}^T\mathbf{W})_{ii} + 2\lambda_1 + 12\lambda_2 (h_{ij}^t)^2 - 12\lambda_2 h_{ij}^t + 2\lambda_2$$

To prove the above inequality, we have

$$\begin{aligned}
\varphi_{ij} h_{ij}^t &= (\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + \lambda_1 h_{ij}^t + 6\lambda_2 (h_{ij}^t)^3 + \lambda_2 h_{ij}^t \\
&= \sum_{l=1}^{k} (\mathbf{W}^T\mathbf{W})_{il} h_{lj}^t + \lambda_1 h_{ij}^t + 6\lambda_2 (h_{ij}^t)^3 + \lambda_2 h_{ij}^t \\
&\geq (\mathbf{W}^T\mathbf{W})_{ii} h_{ij}^t + \lambda_1 h_{ij}^t + 6\lambda_2 (h_{ij}^t)^3 + \lambda_2 h_{ij}^t \\
&\geq h_{ij}^t \big((\mathbf{W}^T\mathbf{W})_{ii} + \lambda_1 + 6\lambda_2 (h_{ij}^t)^2 - 6\lambda_2 h_{ij}^t + \lambda_2 \big) \\
&= \frac{1}{2} F''_{h_{ij}}(h_{ij}^t) h_{ij}^t
\end{aligned}$$

Thus, $G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$.

Replacing $G(h_{ij}, h_{ij}^t)$ in Eq. (4.5) by Eq. (4.6) and setting $\frac{\partial G(h_{ij}, h_{ij}^t)}{\partial h_{ij}}$ to be 0 result in the update rule in Eq. (4.4).

Since Eq. (4.6) is an auxiliary function, $F_{h_{ij}}$ is non-increasing under this up-

date rule.

The update rule for w_{ij} can be proved similarly. □

Our overall approach is described as Algorithm 4.1. **W** and **H** are initialised using plain NMF (Lee and Seung, 1999, 2001), then normalised (Zhang et al., 2007, 2010).

Algorithm 4.1 NMF-Guttman

Require:
$\mathbf{V}$, $\mathbf{H}_{ideal}$, λ_0, λ_1, λ_2, k;
Ensure:
A topic-student matrix, **H**;
1: Initialise **W**, **H** using NMF;
2: Normalise **W**, **H** following (Zhang et al., 2007, 2010);
3: **repeat**
4: Update **W**, **H** iteratively based on Eq. (4.3) and Eq. (4.4);
5: **until** converged
6: **return** **H**;

4.3.3 Selection of $\mathbf{H}_{ideal}$

Topic-student matrix $\mathbf{H}_{ideal}$ is an ideal target where students' topic responses conform to a perfect Guttman scale. $\mathbf{H}_{ideal}$ can be obtained in different ways depending on the attribute of interest to be measured. In this chapter, we are interested in measuring students' latent skill in MOOCs. We envision measurement at the completion of a first offering, with scaled items applied in subsequent offerings for measuring students or curriculum design; alternatively within one offering after a mid-term. Thus, $\mathbf{H}_{ideal}$ can be obtained using assessment, *which need not be based on Guttman-scaled items*. For each student j, his/her responses to the topics given by column $(h_{ideal})_{\cdot j}$ are selected based on his/her grade $g_j \in [0, 100]$, as

$$(h_{ideal})_{\cdot j} = (\underbrace{1 \cdots 1}_{b} \underbrace{0 \cdots 0}_{k-b})$$

$$\text{where} \quad b = \min\left\{ \left\lfloor \frac{g_j + width}{width} \right\rfloor, k \right\}, width = \frac{100}{k} \ .$$

For example, Figure 4.1 shows an exemplar topic-student matrix with ideal Guttman scale. Each column corresponds to a student response pattern on $k = 10$ topics. A student j with $g_j = 35$ is expected to have response pattern $(h_{ideal})_{.j} = (1111000000)$. Similar to the example of a perfect Guttman scale measuring mathematical ability (Table 2.2) in Section 2.3.2, a student who participates a topic also participates topics of lower rank-order. The higher they are graded, the more the topics they participate.

$$\mathbf{H}_{ideal}(\text{example}) \quad \begin{array}{c} \text{grade} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} 8 & 25 & 46 & 67 & 89 & 98 & 78 & 35 & 55 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

Figure 4.1: An exemplar topic-student matrix with ideal Guttman scale

4.4 Experiments

We conduct experiments to evaluate the effectiveness of our algorithms on real MOOCs on Coursera. We also demonstrate the robustness of our approach in terms of parameter sensitivity. In our experiments, we use the first offerings of three Coursera MOOCs from Education, Economics and Computer Science offered by The University of Melbourne. They are *Assessment and Teaching of 21st Century Skills*, *Principles of Macroeconomics*, *Discrete Optimisation* and are named EDU, ECON and OPT for short respectively.

4.4.1 Dataset Preparation

We focus on the students who contributed posts or comments in forums. For each student, we aggregate all posts/comments that s/he contributed. After stemming, removing stop words and html tags, a word-student matrix with normalised tf-idf in [0,1] is produced. The statistics of words and students for the MOOCs are displayed in Table 4.2.

Table 4.2: Statistics of Datasets

MOOC	#Words	#Students
EDU	20,126	1,749
ECON	22,707	1,551
OPT	17,059	1,092

4.4.2 Baseline Approach and Evaluation Metrics

Since there has been no prior method to automatically generate topics forming a Guttman scale, we compare our algorithm with standard NMF (with no regularisation on $\mathbf{H}_{ideal}$).

We adopt the Coefficient of Reproducibility (CR) as it is commonly used to evaluate Guttman scale quality:

$$CR = 1 - \frac{\text{No. of errors}}{\text{No. of possible errors(Total responses)}} \ .$$

CR measures how well a student's responses can be predicted given his/her position on the scale, *i.e.*, total score. By convention, a scale is accepted with items scaled unidimensionally, if its CR is at least 0.90 (Guttman, 1950).

To guarantee binary $\mathbf{H}$, we first scale to $\frac{h_{ij}-\min(\mathbf{H})}{\max(\mathbf{H})-\min(\mathbf{H})} \in [0,1]$, then threshold against a value in $[0.1, 0.2, \cdots, 0.9]$ maximising CR, so that we *conservatively* report CR.

4.4.3 Experimental Setup

Evaluation Setting

We split data into a training set (70% students) and a test set (30% students). The topics are generated by optimising the objective function (5.1) on the training set, and evaluated using CR and the quality of approximation $\|\mathbf{V} - \mathbf{WH}\|_F^2$. To simulate the inferring responses for new students, which has not been explored previously, the trained model is evaluated on the test set using Precision-Recall and ROC curves. Note that in the psychometric literature, validation typically ends with an accepted (> 0.9) CR on the training set.

After learning on the training set word-student matrix $\mathbf{V}^{(train)}$, two matrices are produced: a word-topic matrix $\mathbf{W}^{(train)}$ and topic-student matrix $\mathbf{H}^{(train)}$. To evaluate the trained model on the test set $\mathbf{V}^{(test)}$, we apply the trained word-topic matrix $\mathbf{W}^{(train)}$. Together, we have the relations

$$\begin{aligned} \mathbf{V}^{(train)} &= \mathbf{W}^{*(train)}\mathbf{H}^{(train)} \\ \mathbf{V}^{(test)} &= \mathbf{W}^{*(train)}\mathbf{H}^{(test)} \ . \end{aligned}$$

Solving for $\mathbf{H}^{(test)}$ yields

$$\mathbf{H}^{(test)} = \mathbf{H}^{(train)}(\mathbf{V}^{(train)})^{\dagger}\mathbf{V}^{(test)} \ .$$

where $(\mathbf{V}^{(train)})^{\dagger}$ denotes the pseudoinverse of $\mathbf{V}^{(train)}$.

Hyperparameter Settings

Table 4.3 shows the parameter values used for parameter sensitivity experiments, where the default values in boldface are used in other experiments.

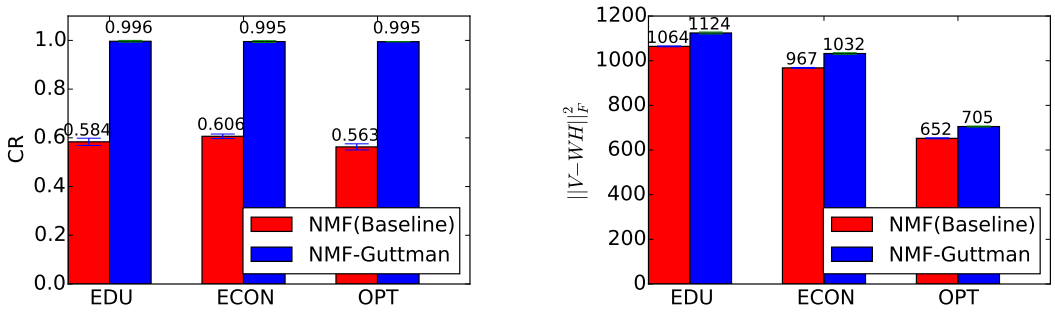
4.4.4 Results

In this group of experiments, we examine how well the generated topics conform to a Guttman scale, and the quality of approximation **WH** to **V**. The reported results are the results averaged over 10 runs. The parameters are set

Table 4.3: Hyperparameter Settings

Parameter	Values Explored (**Default Value**)
λ_0	$[10^{-4}, 10^{-3}, 10^{-2}, \mathbf{10^{-1}}, 10^{0}, 10^{1}, 10^{2}]$
λ_1	$[10^{-4}, 10^{-3}, 10^{-2}, \mathbf{10^{-1}}, 10^{0}, 10^{1}, 10^{2}]$
λ_2	$[10^{-4}, 10^{-3}, \mathbf{10^{-2}}, 10^{-1}, 10^{0}, 10^{1}, 10^{2}]$
k	$[5, \mathbf{10}, 15, 20, 25, 30]$

using the values in boldface in Table 4.3. Figure 4.2 displays the comparison between our algorithm NMF-Guttman and the baseline NMF in terms of CR, and the quality of approximation **WH** to **V** on the training set.



(a) CR

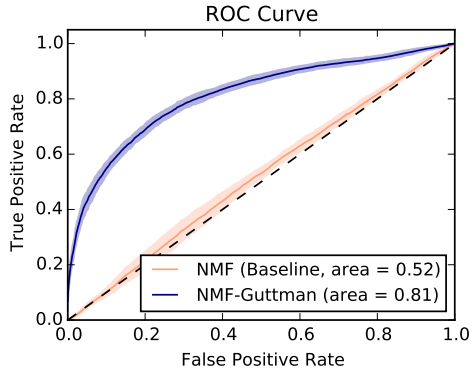
(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$

Figure 4.2: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$.

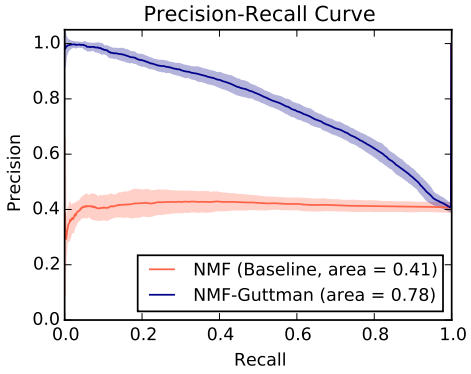
It is clear that our algorithm NMF-Guttman can provide excellent performance in terms of CR with nearly a perfect 1.0, well above the 0.9 cutoff for acceptance. This significantly outperforms baseline which has 0.60 CR across the MOOCs, below Guttman scale acceptance. Meanwhile, NMF-Guttman maintains good quality of approximation, with only slightly inferior $\|\mathbf{V} - \mathbf{WH}\|_F^2$ comparing to NMF (5%, 6%, 8% worse on EDU, ECON, OPT). This is reasonable, as NMF-Guttman has more constraints hence the model itself is less likely to approximate **V** as well as the less constrained standard NMF.

The ROC and Precision-Recall curves (averaged curves with standard deviation over 10 runs) on test set for the ECON MOOC are shown in Figure 4.3. It can be seen that NMF-Guttman significantly dominates NMF, with around

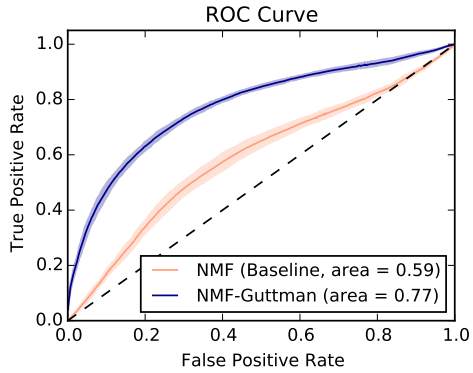
20%-30% better performance, demonstrating the possibility of using the topics for inferring the response of unseen students.



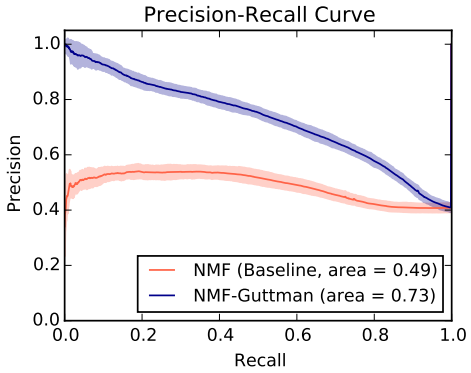
(a) OPT



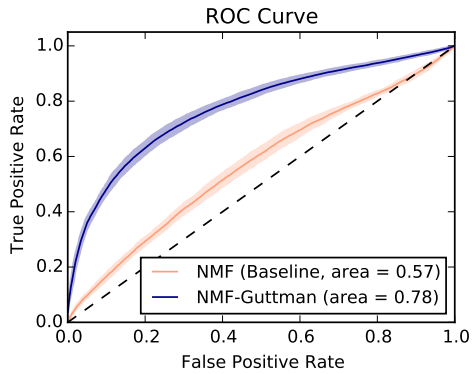
(b) OPT



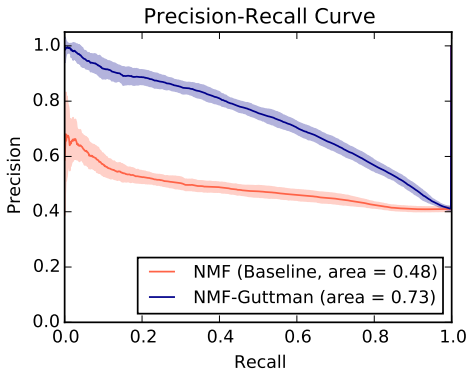
(c) ECON



(d) ECON



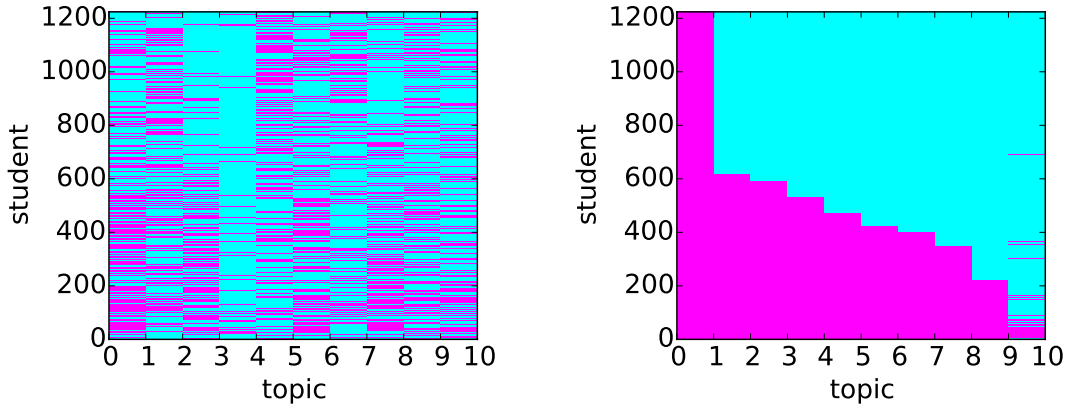
(e) EDU



(f) EDU

Figure 4.3: Comparison of NMF and NMF-Guttman in terms of ROC curve and Precision-Recall curve.

We next visualise the student-topic matrix $\mathbf{H}^T$ produced by NMF and NMF-Guttman respectively. Figure 4.4 is a clear demonstration that NMF-Guttman can produce excellent Guttman scales, while NMF may not. Around half of the cohort (having grade=0) only contribute to topic 1—the easiest—while only a few students contribute to topic 10—the most difficult.



(a) NMF (b) NMF-Guttman

Figure 4.4: Student-topic matrix generated by NMF and NMF-Guttman for MOOC EDU; fuchsia for 1, cyan for 0.

NMF-Guttman can discover items (topics) with responses conforming to a Guttman scale while maintaining the quality of factorisation approximation. It also effectively infers new students' responses.

4.4.5 Validity

The results above establish that our algorithm generates items (topics) with responses conforming to the Guttman scale. Next we test validity—whether topics are meaningful in two aspects: **a) Interpretability**: Are the topics interpretable? **b) Difficulty level**: Do topics exhibit different difficulty levels as inferred by our algorithm and implied by the Guttman scale?

Qualitative Survey

To answer the above questions, we interviewed experts with relevant course background. We showed them the topics (each topic is represented by top 10 words) discovered by our algorithm NMF-Guttman and those generated from the baseline NMF, while blinding interviewees to the topic set's source. We randomised topic orders, and algorithm order, since our algorithm naturally suggests topic order. We posed the following questions for the topics from NMF-Guttman and NMF respectively:

Q1. *Interpretation:* interpret the topic's meaning based on its top 10 word description.

Q2. *Interpretability:* how easy is interpretation? 1=Very difficult; 2=Difficult; 3=Neutral; 4=Easy; 5=Very easy.

Q3. *Difficulty level:* how difficult is the topic to learn? 1=Very easy; 2=Easy; 3=Neutral; 4=Difficult; 5=Very difficult.

Q4. *Ranking:* rank the topics according to their difficulty levels. From 1=easiest; to 10=most difficult.

a) OPT MOOC We interviewed 5 experts with PhDs in discrete optimisation for the OPT MOOC. To validate the topics' difficulty levels, we compute the Spearman's rank correlation coefficient between the ranking from our algorithm and the one from each interviewee, which is shown in Table 4.4. There is high correlation between the NMF-Guttman ranking and those of the Interviewees, *suggesting the topics' Guttman scale relates to difficulty.*

Table 4.4: Survey for OPT MOOC.

Interviewee	Background	Spearman's rank correlation coefficient
1	Works in optimisation and took OPT MOOC	0.71
2	Tutor for OPT MOOC	0.37
3	Professor who teaches optimisation courses	0.90
4	Works in optimisation	0.67
5	Works in optimisation	0.41

Table 4.5 and Table 4.6 show Interviewee 1's interpretation on OPT MOOC topics generated from NMF and NMF-Guttman. It can be seen that the topics from NMF-Guttman are interpretable and exhibit different difficulty levels, qualitatively validating the topics can be used to measure students' latent skill. Note that the topics produced by NMF do not conform to a Guttman scale and are not designed for measurement. Indeed we observed informally that NMF-Guttman's topics were more diverse than those of NMF. For OPT MOOC, half of the topics are not relevant to the course content directly, *i.e.*, feedback about the course and platform/submission issues. While most of the topics from NMF-Guttman are closely relevant to the course content, which are more useful to measure students' skill or conduct curriculum refinement.

b) EDU MOOC Table 4.7 and Table 4.8 show the course coordinator's interpretation on EDU MOOC topics generated from NMF and NMF-Guttman. The course coordinator who has detailed understanding of the course, its curriculum and its forums, was interviewed to answer our survey questions. A 0.8 Spearman's rank correlation coefficient is found between the NMF-Guttman ranking and that of the course coordinator, supporting that the inferred difficulty levels are meaningful. Furthermore, most of the NMF-Guttman's topics are interpretable, less fuzzy, and less overlapping than those of NMF, as judged by the course coordinator.

4.4.6 Parameter Sensitivity

To validate the robustness of parameters and analyse the effect of the parameters, a group of experiments were conducted. The parameter settings are shown in Table 4.3. The performance of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_0, λ_1, λ_2 and k are shown in Figure 4.5 to Figure 4.8.

It can be seen that NMF-Guttman's high performance is stable for λ_1 varying over a wide range 10^{-1} to 10^2. Similar results are found for λ_0, λ_2, and k. NMF-Guttman is not sensitive to λ_0 and k. For λ_2, NMF-Guttman stably performs well when λ_2 varies from 10^{-4} to 10^{-2}. Overall, our algorithm NMF-Guttman is robust, consistently achieves much higher CR than NMF with varying λ_0, λ_1, λ_2 and k, while maintaining the quality of approximation $\|\mathbf{V} - \mathbf{WH}\|_F^2$.

Table 4.5: Interviewee 1's interpretation on OPT MOOC topics generated from NMF.

No.	Topics	Interpretation
1	course thank really learn would like assign time lecture great	Thanks!
2	video lecture load 001 optimization chrome detail coursera org class	Platform
3	file pi line urlib2 lib submit python27 data solver req	Python external solvers
4	submit assignment assignment_id view message screen namehttp 3brows detail 001	Platform/submission
5	problem knapsack solution value optimization submit grade thank solve get	How to submit assignments
6	python solver use matlab command install pyc run java window	Python/Java/Matlab and extend solver (How to start)
7	item value weight capacity estimate take node knapsack tree calculation	Dynamic programing for knapsack, how to understand and code
8	color node graph order clique number use iteration degree edge	Graph coloring, how to use and understand graph theory concepts
9	solution opt problem use search get custom move time optimize	Traveling salesman problem, trying to improve algorithm/customise
10	use dp memory column bb algorithm bound table implement time	Comparing algorithms memory/time

Table 4.6: Interviewee 1's interpretation on OPT MOOC topics generated from NMF-Guttman with inferred difficulty ranking.

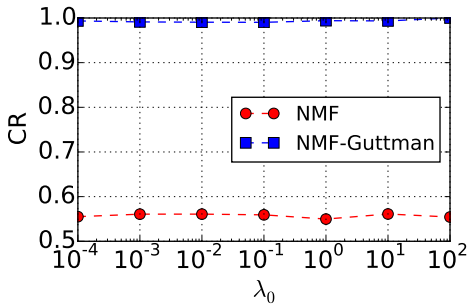
No.	Topics	Interpretation	Inferred Ranking
1	python problem file solver assign pi class video course use	How to use platform/python	1
2	submit thank please pyc grade feedback run solution check object	Platform/submission issues	2
3	warehouse one result edge exactly list decide tour lib suppose	How to read&use data for facility location	3
4	solution optimize best first much insert move want feasible less	How to improve/create heuristics for knapsack problems	4
5	color opt random search local greedy swap node good get	Understand and implement local search	5
6	point mip certificate puzzle enough le route model course de	Course structure (eg. what's enough to get certificate?)	6
7	use scip two try implement lp differ need solver easy	How to implement LP/MIP and solver availability	7
8	time temperature sa move opt would like well start ls	How to design and tune simulated annealing and local search	8
9	problem warehouse custom 10 tsp cluster mip constraint vehicle solution	Relationship between problems and algorithms	9
10	item use value node solution problem algorithm optimize time dp	Knapsack (using dynamic programing), how to speed up algorithms	10

Table 4.7: The course coordinator's interpretation on EDU MOOC topics generated from NMF.

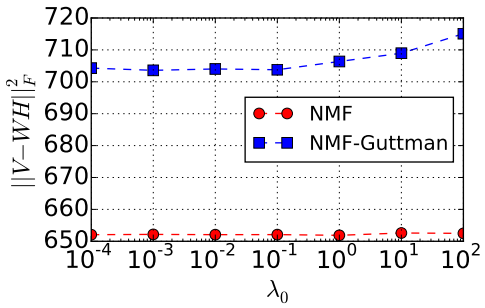
No.	Topics	Interpretation
1	student cp think use would skill teacher need task assess	Discussion of assesemnt of collaborative problems solving discussed in week 2 of the course
2	teach course hello teacher english hope everyone name hi improve	Welcome introductions to the course
3	assign peer evaluate grade thank course score mooc mark assess	Discussion about peer assessemtns in the course
4	skill century 21st assess develop learn curriculum need interest education	General discussion of introducory ideas in the course
5	org 001 atc21s http coursera 971791 human_grading assessments courses class	General discussion about the approach to assessemtns in the course
6	problem solve collaborate idea skill think differ group task cp	Discussion about the nature of collaborative problems solving, week 2 of the course
7	learn forward look student assess hi excit collaborate everyone course	Introductory comments about the course
8	technology learn use education teacher new learner change us way	Discussion about the impact of technology on the curriculum
9	school education year australia current primary hi interest melbourne name	Talk between participants about their background
10	work thank group really help together routine know time student	Thank you notes and discussion at the end of the course

Table 4.8: The course coordinator's interpretation on EDU MOOC topics generated from NMF-Guttman with inferred difficulty ranking.

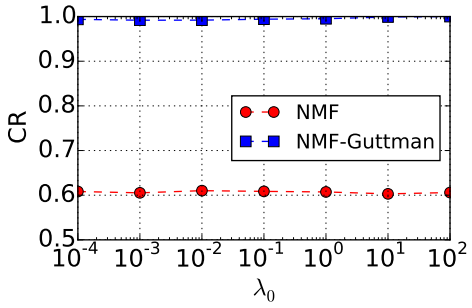
No.	Topics	Interpretation	Inferred Ranking
1	learn student teach teacher skill course assess use school collaboration	Discussion of relationship between teaching and learning as in week 1 of the course syllabus	1
2	forward name also philippin hi join india teach help better	Establishing social presence: Introduction posts, and statements of and what people want to get out of the course	2
3	assign peer thank one evaluate mark grade comment score could	Discussion about the peer assignments in the course	3
4	skill develop need assess australia plan base progress social approach	Discussion of developmental teaching and assessment as per 1st and 2nd week of course syllabus	4
5	001 coursera org atc21s 971791 human_grading courses assessments submissions class	General discussion about course process and structure	5
6	cp task problem think collaborate solve differ activity group idea	Discussion of collaborative problem solving, which is the focus of week 2 syllabus	6
7	student assess individual australia provide report less level progress may	Discussion of student level, individualised reporting against performance levels, focus of weeks 2 and 3 of syllabus	7
8	use make great give many way thank bring agree human	Appreciation posts at the end of the course	8
9	would school level week link read model table set observe	Unclear	9
10	work student group thank one need time make together know	Discussion of difference between collaboration and groupwork, theme through the course, and in assignments	10



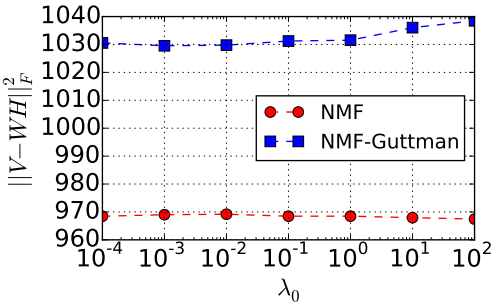
(a) CR on OPT



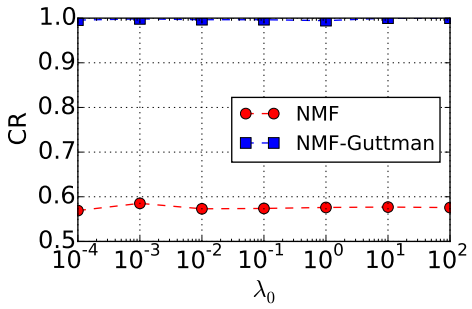
(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



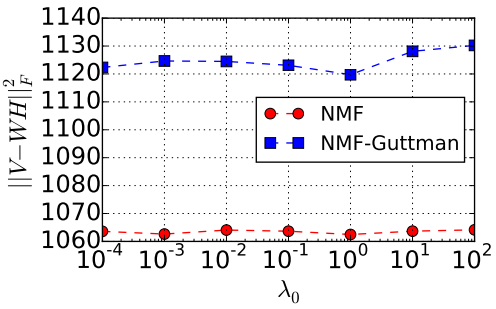
(c) CR on ECON



(d) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON

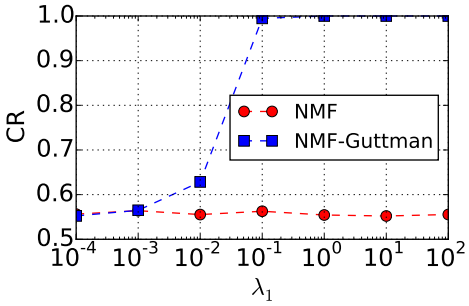


(e) CR on EDU

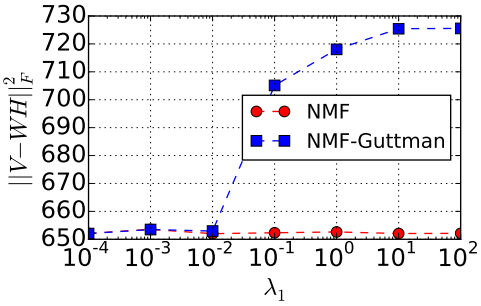


(f) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU

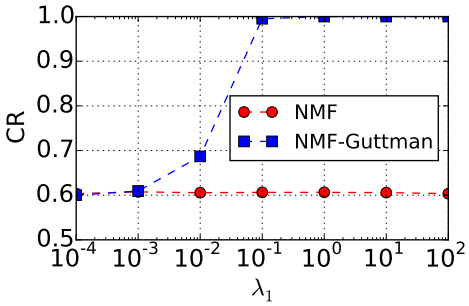
Figure 4.5: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_0.



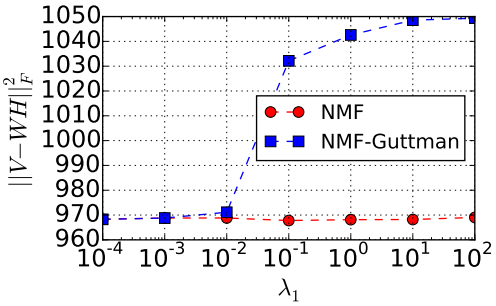
(a) CR on OPT



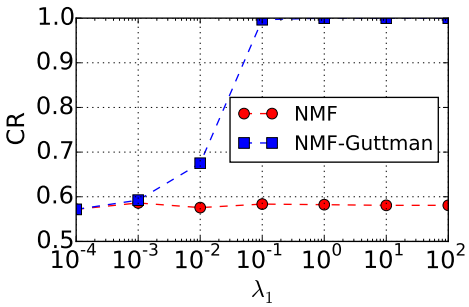
(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



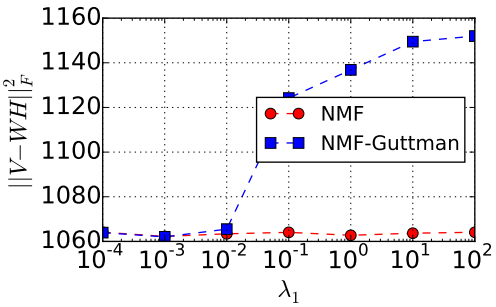
(c) CR on ECON



(d) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON

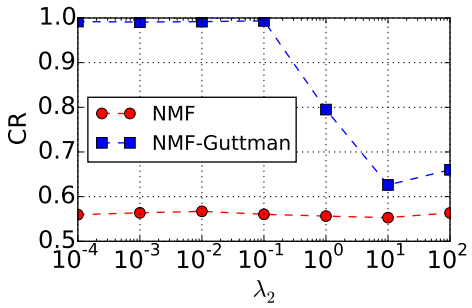


(e) CR on EDU

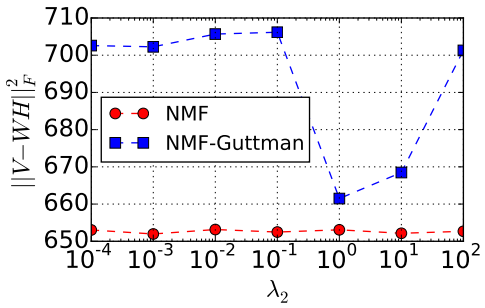


(f) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU

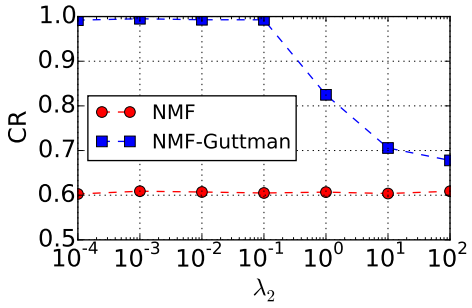
Figure 4.6: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_1.



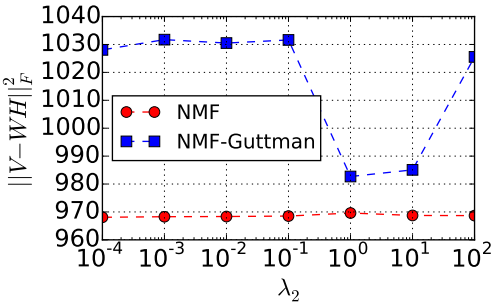
(a) CR on OPT



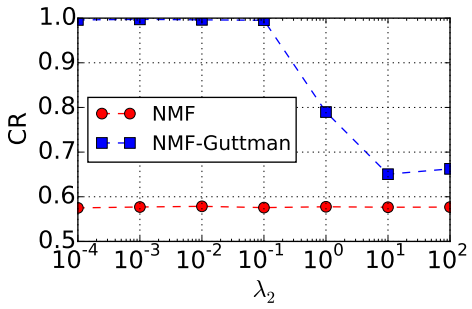
(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



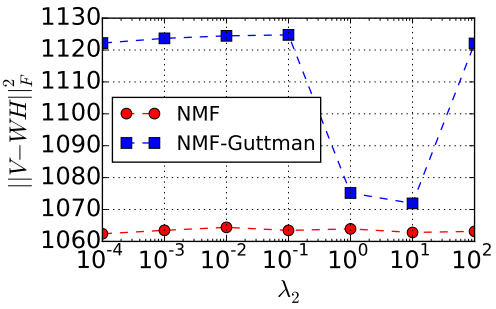
(c) CR on ECON



(d) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON

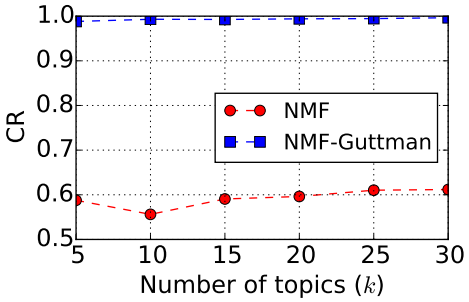


(e) CR on EDU

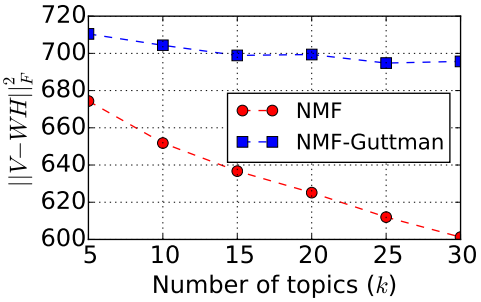


(f) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU

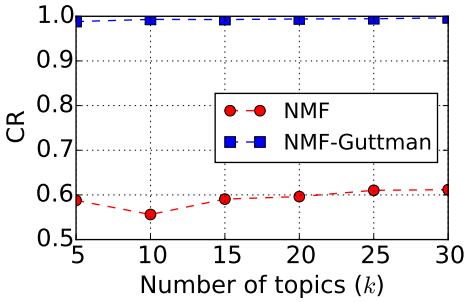
Figure 4.7: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying λ_2.



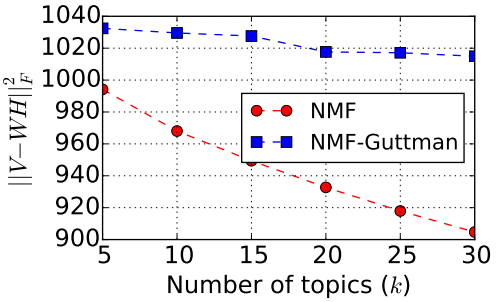
(a) CR on OPT



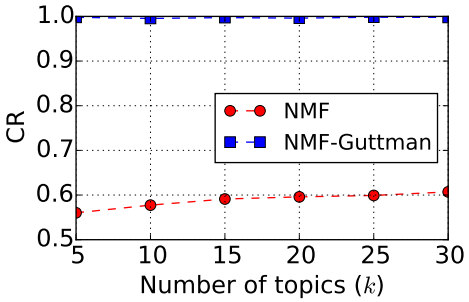
(b) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



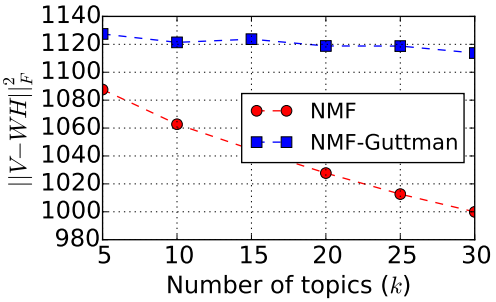
(c) CR on ECON



(d) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON



(e) CR on EDU



(f) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU

Figure 4.8: Comparison of NMF and NMF-Guttman in terms of CR and $\|\mathbf{V} - \mathbf{WH}\|_F^2$ with varying the number of topics (k).

4.5 Conclusion

This is the first study that combines a machine learning technique (topic modelling) with measurement theory (psychometrics) as used in education. Inspired by the recent research on measurement of student learning in the education community and the importance of forum discussion for understanding student learning and providing feedback, we explore the suitability of using students' participation across automatic discovered topics as evidence of measuring student ability considering the Guttman scale. To establish the statistical effectiveness, the topics are required to conform to the Guttman scale as evidence of measuring educationally-meaningful skill attainment. We achieve this by a novel regularisation of non-negative matrix factorisation.

Our empirical results are compelling and extensive. We contribute a quantitative validation on three Coursera MOOCs, demonstrating our algorithm conforms to Guttman scaling (shown with high coefficients of reproducibility), strong quality of factorisation approximation, and predictive power on unseen students (via ROC and PR curve analysis). We also contribute a qualitative study of domain expert interpretations on two MOOCs, showing that most of the topics with difficulty levels inferred, are interpretable and meaningful.

Chapter 5

Topic-Instrumented Measurement Based on the Rasch Model

Continuing the theme of Chapter 4, this chapter again explores the problem of the suitability of using automatically discovered topics from MOOC discussion forum content for modelling students' academic abilities. However, we now investigate the use of a different measurement model, the Rasch model from Item Response Theory to evaluate the statistical effectiveness. The Rasch model brings several advantages over the Guttman scale, enabling adaptive and personalised feedback. The algorithm we previously proposed for the Guttman scale in Chapter 4 cannot be used or adapted readily for Rasch modelling, as the Rasch model is more statistically and computationally complex. In this chapter, we propose the TopicResponse algorithm, which simultaneously performs topic modelling and Rasch model fitting, to automatically discover topics such that students' participation across them fits the Rasch model. This chapter is based on the following manuscript: Jiazhen He, Rui Zhang, James Bailey, Benjamin I.P. Rubinstein, and Sandra Milligan. "TopicResponse: A Marriage of Topic Modelling and Rasch Modelling for Automatic Measurement in MOOCs". Under second round major revision for *Machine Learning Journal*, 2016.

5.1 Introduction

In Chapter 4, we have adapted NMF-based topic modelling to the psychometric testing of MOOC students based on their online forum postings under the measurement model of Guttman scale. However, the Guttman scale is regarded to be an overly-idealised model, which is impractical in the real world. In contrast the Rasch model, one of the simplest item response theory (IRT) model and the basis for many extensions, has been widely used in education and psychology. It is a generative probabilistic model that represents student responses as a noisy observation of latent student abilities related to item difficulties. It can be viewed as a stochastic counterpart to the Guttman scale, permitting measurement error. If a person's ability level is higher than an item's difficulty, the person will answer the item correctly in the Guttman scale, while in the Rasch model, there is a certain probability of incorrect response. While the Guttman scale only permits ordering of persons and items on a latent scale, Rasch permits the meaningful interpretation of the differences between them (Scholten, 2011).

We investigate whether students' participation in automatically discovered forum topics can be used as an instrument to model students' ability. If students' participation across the discovered topics fit the Rasch model in terms of statistical effectiveness, and the topics are interpretable to subject-matter experts by way of qualitative effectiveness, then the discovered topics can be regarded as useful items for measurement. The resulting scaled topics, endowed with estimated difficulty levels, can assist in subsequent curriculum refinement, student assessment, and personalised feedback.

The technical challenge is to automatically discover topics such that students' participation across them fit the Rasch model. The algorithm proposed for the Guttman scale in Chapter 4 does not adapt readily for Rasch modelling. Instead we propose the TopicResponse algorithm, which simultaneously performs non-negative matrix factorisation and Rasch model fitting. The main contributions of this chapter include:

- The first study that combines topic modelling with Rasch modelling in psychometric testing: generating topics that measure students' academic

abilities based on online forum postings;

- An algorithm combining NMF and Rasch model (as evidence of educationally meaningful ability); and

- Quantitative experiments on three Coursera MOOCs covering a broad swath of disciplines, establishing statistical effectiveness of our algorithm, and qualitative results on a Discrete Optimisation MOOC, supporting interpretability.

5.2 Preliminaries and Problem Formulation

We firstly give a brief overview of the Rasch model for dichotomous data and its estimation, and then define our problem. Note that more detailed introduction about the Rasch model is given in Section 2.4.1.

As introduced in Section 2.4.1, the Rasch model for dichotomous data (correct/incorrect, agree/disagree responses) specifies the probability of a person's positive response (correct, agree) on an item as a logistic function of the difference between the person's ability and item difficulty, which can be formalised as

$$p_{ij} = P(X_{ij} = 1|\beta_i, \theta_j) = \frac{1}{1 + \exp\left(-\left(\theta_j - \beta_i\right)\right)} \ , \tag{5.1}$$

where θ_j denotes person j's ability, β_i denotes item i's difficulty, X_{ij} denotes the person j's response on item i, and p_{ij} denotes the probability of person j's positive response on item i.

Given an observed response matrix $\mathbf{x}$=$[x_{ij}]$ (*e.g.*, Table 2.3), the goal is to estimate the person and item parameters θ_j and β_i, which can be estimated by joint maximum likelihood using the iterative Newton-Raphson method, with the logarithm of the likelihood function to be maximised as follow

$$\log \mathcal{L}(\boldsymbol{\beta}, \boldsymbol{\theta}|\mathbf{x}) = \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}(\theta_j - \beta_i) - \sum_{i=1}^{k}\sum_{j=1}^{n} \log(1 + \exp\left(\theta_j - \beta_i\right)) \ . \tag{5.2}$$

5.2.1 Problem Formulation

We seek to explore the feasibility of automatic discovery of forum discussion topics for measuring students' academic abilities in MOOCs, as quantified by the Rasch model. Our central tenet is that topics can be regarded as useful items for measuring a latent skill, if student responses to these topics are well fit by the Rasch model, and if the topics are interpretable to domain experts for educational relevance. Therefore, we need to discover topics from students' posts and comments in MOOC forums, in such a way that students' participation across these topics fits the Rasch model. Student item response records whether a student posts on the corresponding topic or not. After discovery, topics must then be further assessed for interpretability to domain experts. Our goal is decision support.

In particular, under the NMF framework, a word-student matrix $\mathbf{V}$ can be factorised into two non-negative matrices: word-topic matrix $\mathbf{W}$ and topic-student matrix $\mathbf{H}$. Our application requires that the topic-student matrix $\mathbf{H}$ be **a)** binary ensuring the response of a student to a topic is dichotomous; **b)** useful for measuring students' academic abilities; and **c)** well-fit by the Rasch model. NMF provides an elegant framework for incorporating these constraints via adding novel regularisation, as detailed in the next section. A glossary of the symbols most used in this chapter is given in Table 5.1.

5.3 TOPICRESPONSE Model: Joint NMF-Rasch Model

To favour the topics which fit the Rasch model, we jointly optimise the NMF and the Rasch model, which yields the objective function

$$g(\mathbf{W}, \mathbf{H}, \theta, \beta) = \|\mathbf{V} - \mathbf{WH}\|_F^2 - \lambda_0 f_R(\mathbf{H}, \theta, \beta) \ ,$$

$$f_R(\mathbf{H}, \theta, \beta) = \sum_{i=1}^{k}\sum_{j=1}^{n} h_{ij}(\theta_j - \beta_i) - \sum_{i=1}^{k}\sum_{j=1}^{n} \log\left(1 + \exp\left(\theta_j - \beta_i\right)\right) ,$$

where $f_R(\mathbf{H}, \theta, \beta)$ is the log likelihood function to be maximised in Rasch estimation as introduced in Section 2.4.1, and $\lambda_0 > 0$ is a user-specified parameter

Table 5.1: Glossary of symbols

Symbol	Description
m	the number of words
n	the number of students
k	the number of topics
$\mathbf{V} = (v_{ij})_{m\times n}$	word-student matrix
$\mathbf{W} = (w_{ij})_{m\times k}$	word-topic matrix
$\mathbf{H} = (h_{ij})_{k\times n}$	topic-student matrix
$\mathbf{H}_{ideal} = \left((h_{ideal})_{ij}\right)_{1\times n}$	matrix for students with ideal total number of topics posted
$\mathbf{1}_r$	all-ones matrix with size $1 \times n$
g_j	student j's grade
$\boldsymbol{\beta} = (\beta_i)_k$	item difficulty vector
$\boldsymbol{\theta} = (\theta_j)_n$	student ability vector
p_{ij}	the probability of positive response of person j to item i
$\lambda_0, \lambda_1, \lambda_2, \lambda_3$	regularisation coefficients

controlling the trade-off between the quality of factorisation and Rasch estimation.

Weak supervision of item responses. The fit between student topic responses $\mathbf{H}$ and the Rasch model will provide statistical evidence of measuring skill attainment. However, it is difficult to conclude what the topics are measuring without domain knowledge. To favour the topics that can be used to measure students' academic abilities, we impose a constraint on $\mathbf{H}$ based on some student grade, which provides an indicator of student's abilities (we discuss sources of auxiliary grade information below). In particular, we assume that there is the following relationship between the ideal number of distinct topics that each student j contributes and their grade $g_j \in [0, 100]$,

$$(h_{ideal})_{1j} = \min\left\{ \left\lfloor \frac{g_j + width}{width} \right\rfloor, k-1 \right\} \ , \ width = \frac{100}{k-1} \ ,$$

where $\mathbf{H}_{ideal}$ is a $1 \times n$ matrix, denoting the ideal number of distinct topics posted by students. For example under $k = 10$ items, student j scoring $g_j = 45$ should post on a number of topics $(h_{ideal})_{1j} = 5$. The minimum and maximum

$$\mathbf{H}_{ideal} \text{ for Guttman scale} = \begin{array}{c} \text{grade} \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} 8 & 25 & 46 & 67 & 89 & 98 & 78 & 35 & 55 \end{array} \\ \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \end{pmatrix} \end{array}$$

$$\mathbf{H}_{ideal} \text{ for Rasch model} = \begin{array}{c} \text{grade} \\ \\ \end{array} \begin{array}{c} \begin{array}{ccccccccc} 8 & 25 & 46 & 67 & 89 & 98 & 78 & 35 & 55 \end{array} \\ \begin{pmatrix} 1 & 3 & 5 & 7 & 9 & 9 & 8 & 4 & 5 \end{pmatrix} \end{array}$$

Figure 5.1: An exemplar of $\mathbf{H}_{ideal}$ in the Guttman scale and the Rasch model.

number of different topics that a student j posted is 1 and $k-1$ respectively. This is motivated by the initialisation of θ and β as illustrated in Section 2.4.1, where positive responses on 0 or k topics is undesirable.

This supervision constraint is markedly weaker than a similar constraint found in (He et al., 2016), as demonstrated in Figure 5.1. He et al. (2016) leverage the student grade to exactly determine the item responses for the Guttman scale. The Guttman scale, as a deterministic model, requires that if a student can get a difficult item correct, they can also achieve correct responses on all easier items. This assumption is very restrictive, and rarely makes sense in practice. The Rasch model allows errors in the responses; and only constrains the number of distinct topics posted by a student, rather than the exact response pattern.

Most (MOOC) courses conduct multiple forms of assessment throughout the duration of teaching. For example, weekly quizzes, take-home assignments, mid-term tests, projects, presentations, etc. In the large-scale MOOC context, such evaluations may be peer-assessed. Students often enter courses with some cumulative grade-point average that may be (loosely) predictive of future performance. Any of these readily-available sources of student information could

be reasonably used to seed $\mathbf{H}_{ideal}$. Even final course grades could be used, particularly when the ultimate application of TopicResponse is not measuring students, but refining curriculum.

In order to encourage satisfaction of the $\mathbf{H}_{ideal}$ soft constraint on topic responses, we introduce a regularisation term on $\mathbf{H}$, namely $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|$.

Quantising & Regularising the Response Matrix. We introduce regularisation term $\|\mathbf{W}\|$, commonly used to prevent overfitting in NMF. To encourage binary solutions, we impose an additional regularisation term $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|$, where operator $\circ$ denotes the Hadamard product. Binary matrix factorisation (BMF) is a variation of NMF, where the input matrix and the two factorised matrices are all binary. Our approach is inspired by those of Zhang et al. (2007) and Zhang et al. (2010). Our added term equals $\|\mathbf{H} \circ (\mathbf{H} - \mathbf{1})\|$, which is minimised by (only) binary $\mathbf{H}$.

TopicResponse Model. We have the following regularisations:

- $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|$ to encourage a grade-guided $\mathbf{H}$;
- $\|\mathbf{W}\|$ to prevent overfitting; and
- $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|$ to encourage a binary item-response solution.

These terms together with joint NMF-Rasch estimation yield final objective

$$\begin{aligned} f(\mathbf{W}, \mathbf{H}, \theta, \boldsymbol{\beta}) = \|\mathbf{V} - \mathbf{W}\mathbf{H}\| - \lambda_0 f_R(\mathbf{H}, \theta, \boldsymbol{\beta}) + \lambda_1 \|\mathbf{W}\| \\ + \lambda_2 \|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\| + \lambda_3 \|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\| \quad , \end{aligned} \tag{5.3}$$

where $\lambda_1, \lambda_2, \lambda_3 > 0$ are user-specified regularisation parameters, with primal program

$$\underset{\mathbf{W},\mathbf{H},\theta,\boldsymbol{\beta}}{\operatorname{argmin}} f(\mathbf{W}, \mathbf{H}, \theta, \boldsymbol{\beta}) \quad \text{s.t.} \quad \mathbf{W} \geq \mathbf{0},\ \mathbf{H} \geq \mathbf{0} \ . \tag{5.4}$$

TopicResponse Fitting Procedure. A local optimum of program (5.4) is achieved via iteration

$$w_{ij} \leftarrow w_{ij}\frac{(\mathbf{V}\mathbf{H}^T)_{ij}}{(\mathbf{W}\mathbf{H}\mathbf{H}^T+\lambda_0\mathbf{W})_{ij}} \tag{5.5}$$

$$h_{ij} \leftarrow h_{ij}\frac{2(\mathbf{W}^T\mathbf{V})_{ij}+8\lambda_2 h_{ij}^3+6\lambda_2 h_{ij}^2+2\lambda_1(\mathbf{1}_r^T\mathbf{H}_{ideal})_{ij}+\lambda_3(\theta-\beta)_{ij}^+}{2(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij}+12\lambda_2 h_{ij}^3+2\lambda_1(\mathbf{1}_r\mathbf{H})_{ij}+2\lambda_2 h_{ij}+\lambda_3(\theta-\beta)_{ij}^-} \tag{5.6}$$

$$\beta_i \leftarrow \beta_i - \frac{\sum_{j=1}^n(p_{ij}-h_{ij})}{-\sum_{j=1}^n p_{ij}(1-p_{ij})} \tag{5.7}$$

$$\theta_j \leftarrow \theta_j - \frac{\sum_{i=1}^k(h_{ij}-p_{ij})}{-\sum_{i=1}^k p_{ij}(1-p_{ij})} \tag{5.8}$$

where

$$\begin{aligned}(\theta-\beta) &= (\theta-\beta)^+-(\theta-\beta)^-\\ (\theta-\beta)_{ij}^+ &= \begin{cases}(\theta-\beta)_{ij} & \text{if } (\theta-\beta)_{ij}>0\\ 0 & \text{if otherwise}\end{cases}\\ (\theta-\beta)_{ij}^- &= \begin{cases}-(\theta-\beta)_{ij} & \text{if } (\theta-\beta)_{ij}<0\\ 0 & \text{if otherwise}\end{cases}\end{aligned}$$

$(\theta-\beta)^+$ and $(\theta-\beta)^-$ denote the positive part and negative part of matrix $(\theta-\beta)$ respectively. We next describe how these update rules are derived.

The update rules (5.7) and (5.8) can be obtained using Newton's method. The update rules (5.5) and (5.6) can be derived via the Karush-Kuhn-Tucker conditions necessary for local optimality. First we construct the unconstrained Lagrangian

$$\mathcal{L}(\mathbf{W},\mathbf{H},\theta,\beta,\boldsymbol{\alpha},\boldsymbol{\gamma}) = f(\mathbf{W},\mathbf{H},\theta,\beta)+\mathrm{tr}(\boldsymbol{\alpha}\mathbf{W})+\mathrm{tr}(\boldsymbol{\gamma}\mathbf{H})\ ,$$

where $\alpha_{ij},\gamma_{ij}\leq 0$ are the Lagrangian dual variables for inequality constraints $w_{ij}\geq 0$ and $h_{ij}\geq 0$ respectively, and $\boldsymbol{\alpha}=[\alpha_{ij}]$, $\boldsymbol{\gamma}=[\gamma_{ij}]$ denote their corresponding matrices. The KKT condition of stationarity requires that the deriva-

tive of $\mathcal{L}$ with respect to $\mathbf{H}$, vanishes at a local optimum $\mathbf{H}^\star, \mathbf{W}^\star, \boldsymbol{\alpha}^\star, \boldsymbol{\gamma}^\star$:

$$\begin{aligned}
\frac{\partial \mathcal{L}}{\partial \mathbf{W}} =& 2\left(\mathbf{W}^\star\mathbf{H}^\star\mathbf{H}^{\star T} - \mathbf{V}\mathbf{H}^{\star T} + \lambda_0\mathbf{W}^\star\right) + \boldsymbol{\alpha}^\star = \mathbf{0} \ , \\
\frac{\partial \mathcal{L}}{\partial \mathbf{H}} =& 2\Big(\mathbf{W}^{\star T}\mathbf{W}^\star\mathbf{H}^\star - \mathbf{W}^{\star T}\mathbf{V} + \lambda_1\mathbf{1}_r^T\mathbf{1}_r\mathbf{H} + \lambda_2\mathbf{H}^\star - \lambda_1\mathbf{1}_r^T\mathbf{H}_{ideal}\Big) \\
& + 4\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - 6\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star - \lambda_3\left((\boldsymbol{\theta} - \boldsymbol{\beta})^+ - (\boldsymbol{\theta} - \boldsymbol{\beta})^-\right) + \boldsymbol{\gamma}^\star \\
=& \mathbf{0} \ .
\end{aligned}$$

Complementary slackness $\gamma_{ij}^\star h_{ij}^\star = 0$, implies:

$$\begin{aligned}
0 &= \left(\mathbf{V}\mathbf{H}^{\star T} - \mathbf{W}^\star\mathbf{H}^\star\mathbf{H}^{\star T} - \lambda_0\mathbf{W}^\star\right)_{ij} w_{ij}^\star \ , \\
0 &= \Big(2\mathbf{W}^{\star T}\mathbf{V} + 6\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star + 2\lambda_1\mathbf{1}_r^T\mathbf{H}_{ideal} - 2\mathbf{W}^{\star T}\mathbf{W}^\star\mathbf{H}^\star - 2\lambda_1\mathbf{1}_r^T\mathbf{1}_r\mathbf{H}^\star \\
&\quad -4\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - 2\lambda_2\mathbf{H}^\star + \lambda_3(\boldsymbol{\theta} - \boldsymbol{\beta})^+ - \lambda_3(\boldsymbol{\theta} - \boldsymbol{\beta})^- \\
&\quad +8\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star - 8\lambda_2\mathbf{H}^\star \circ \mathbf{H}^\star \circ \mathbf{H}^\star\Big)_{ij} h_{ij}^\star \ .
\end{aligned}$$

These two equations lead to the updating rules (5.5) and (5.6). Regarding the update rules (5.5), (5.6), (5.7) and (5.8) we have the following theorem:

Theorem 5.1. *The objective function $f(\mathbf{W}, \mathbf{H}, \boldsymbol{\theta}, \boldsymbol{\beta})$ of TopicResponse program* (5.4) *is non-increasing under update rules* (5.5), (5.6), (5.7) *and* (5.8).

Proof. The update rules for β_i and θ_j are derived using Newton-Raphson method, where the convergence to a local optimum is guaranteed. Here, we focus on the proof for the update rule for h_{ij}. The update rule for w_{ij} can be proved similarly. We follow the similar procedure described in (Lee and Seung, 2001), where an auxiliary function similar to that used in the Expectation-Maximization (EM) algorithm is used for proof.

Definition 5.2. *(Lee and Seung, 2001) $G(h, h')$ is an auxiliary function for $F(h)$ if the conditions*

$$G(h, h') \geq F(h), \qquad G(h, h) = F(h)$$

are satisfied.

Lemma 3. *(Lee and Seung, 2001) If G is an auxiliary function, then F is non-increasing under the update*

$$h_{t+1} = \underset{h}{\operatorname{argmin}}\, G(h, h^t) \tag{5.9}$$

Proof: $F(h^{t+1}) \leq G(h^{t+1}, h^t) \leq G(h^t, h^t) = F(h^t)$

For any element h_{ij} in $\mathbf{H}$, let $F_{h_{ij}}$ denote the part of $f(\mathbf{W}, \mathbf{H}, \theta, \boldsymbol{\beta})$ in Eq. (5.3) in the chapter relevant to h_{ij}. Since the update is essentially element-wise, it is sufficient to show that each $F_{h_{ij}}$ is non-increasing under the update rule of Eq. (5.6). To prove it, we define the auxiliary function regarding h_{ij} as follows.

Lemma 4. *Function*

$$G(h_{ij}, h_{ij}^t) = F_{h_{ij}}(h_{ij}^t) + F'_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t) + \varphi_{ij}(h_{ij} - h_{ij}^t)^2 \tag{5.10}$$

where

$$\varphi_{ij} = \frac{2(\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + 2\lambda_1(\mathbf{1}_r^T\mathbf{1}_r\mathbf{H})_{ij} + 12\lambda_2(h_{ij}^t)^3 + 2\lambda_2 h_{ij}^t + \lambda_3(\theta - \boldsymbol{\beta})_{ij}^-}{2h_{ij}^t}$$

is an auxiliary function for $F_{h_{ij}}$.

Proof: It is obvious that $G(h_{ij}, h_{ij}) = F_{h_{ij}}$. So we only need to prove that $G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$. Considering the Taylor series expansion of $F_{h_{ij}}$,

$$F_{h_{ij}} = F_{h_{ij}}(h_{ij}^t) + F'_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t) + \frac{1}{2}F''_{h_{ij}}(h_{ij}^t)(h_{ij} - h_{ij}^t)^2$$

$G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$ is equivalent to $\varphi_{ij} \geq \frac{1}{2}F''_{h_{ij}}(h_{ij}^t)$, where

$$F''_{h_{ij}}(h_{ij}^t) = 2(\mathbf{W}^T\mathbf{W})_{ii} + 2\lambda_1(\mathbf{1}_r^T\mathbf{1}_r)ii + 12\lambda_2(h_{ij}^t)^2 - 12\lambda_2 h_{ij}^t + 2\lambda_2$$

To prove the above inequality, we have

$$\begin{aligned}
\varphi_{ij}h_{ij}^t &= (\mathbf{W}^T\mathbf{W}\mathbf{H})_{ij} + \lambda_1(\mathbf{1}_r^T\mathbf{1}_r\mathbf{H})_{ij} + 6\lambda_2(h_{ij}^t)^3 + \lambda_2 h_{ij}^t + 0.5\lambda_3(\theta - \beta)_{ij}^- \\
&= \sum_{l=1}^{k}(\mathbf{W}^T\mathbf{W})_{il}h_{lj}^t + \lambda_1\sum_{l=1}^{k}(\mathbf{1}_r^T\mathbf{1}_r)_{il}h_{lj}^t + +6\lambda_2(h_{ij}^t)^3 + \lambda_2 h_{ij}^t + 0.5\lambda_3(\theta - \beta)_{ij}^- \\
&\geq (\mathbf{W}^T\mathbf{W})_{ii}h_{ij}^t + \lambda_1(\mathbf{1}_r^T\mathbf{1}_r)_{ii}h_{ij}^t + 6\lambda_2(h_{ij}^t)^3 + \lambda_2 h_{ij}^t - 12\lambda_2 h_{ij}^t \\
&\geq h_{ij}^t\big((\mathbf{W}^T\mathbf{W})_{ii} + \lambda_1(\mathbf{1}_r^T\mathbf{1}_r)_{ii} + 6\lambda_2(h_{ij}^t)^2 - 6\lambda_2 h_{ij}^t + \lambda_2\big) \\
&= \frac{1}{2}F''_{h_{ij}}(h_{ij}^t)h_{ij}^t
\end{aligned}$$

Thus, $G(h_{ij}, h_{ij}^t) \geq F_{h_{ij}}$.

Replacing $G(h_{ij}, h_{ij}^t)$ in Eq. (5.9) by Eq. (5.10) and setting $\frac{\partial G(h_{ij}, h_{ij}^t)}{\partial h_{ij}}$ to be 0 result in the update rule in Eq. (5.6) in the chapter. Since Eq. (5.10) is an auxiliary function, $F_{h_{ij}}$ is non-increasing under this update rule. □

Algorithm. Our overall approach TopicResponse is described as Algorithm 5.1. $\mathbf{W}$ and $\mathbf{H}$ are initialised using plain NMF (Lee and Seung, 1999, 2001), then normalised (Zhang et al., 2007, 2010). θ and β are initialised based on Eq. (2.7) and Eq. (2.8), where x_{ij} is replaced by h_{ij}. At optimisation completion, estimates for topics, item difficulties and person abilities can be obtained together.

Algorithm 5.1 TopicResponse

Require:
$\mathbf{V}$, $\mathbf{H}_{ideal}$, λ_0, λ_1, λ_2, λ_3,k;
Ensure:
A topic-student matrix, $\mathbf{H}$, item difficulties β, person abilities θ;
1: Initialise $\mathbf{W}$, $\mathbf{H}$ using NMF;
2: Normalise $\mathbf{W}$, $\mathbf{H}$ following (Zhang et al., 2007, 2010);
3: Initialise θ, β based on Eq. (2.7) and Eq. (2.8);
4: **repeat**
5: Update $\mathbf{W}, \mathbf{H}, \beta, \theta$ iteratively based on Eq. (5.5) to Eq. (5.8);
6: **until** converged
7: **return** $\mathbf{H}$;

5.4 Experiments

We use the same datasets in Chapter 4 to evaluate the effectiveness of our algorithms.

5.4.1 Baseline and Evaluation Metrics

We compare our algorithm TopicResponse with the baseline algorithm Grade-Guided NMF (GG-NMF), with the following objection function minimised

$$f_G(\mathbf{W}, \mathbf{H}) = \|\mathbf{V} - \mathbf{WH}\|_F^2 + \lambda_1 \|\mathbf{W}\|_F^2 + \lambda_2 \|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2 + \lambda_3 \|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2 \ .$$

A local optimum can be obtained using the Karush-Kuhn-Tucker conditions. Like TopicResponse, GG-NMF regularises $\mathbf{H}$ by considering the students' grades as an indicator of academic ability. The difference is that TopicResponse optimises the Rasch estimation and NMF simulationeouly, while in GG-NMF, the students' topic responses $\mathbf{H}$ are first obtained, and then are passed through the Rasch model. We evaluate the two algorithms in terms of the following metrics.

- *Quality of factorisation (topics)*, $\|\mathbf{V} - \mathbf{WH}\|_F^2$.
- *Measuring student academic ability instead of other ability or skill*. Quality of constraint on students' topic participation based on their grades, $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$, an indicator of academic ability.
- *Negative log-likelihood.* Log-likelihood implies the fit of the whole data to the Rasch model. For convenience, we look at the negative log-likelihood, which is minimised. Smaller is better. This is our main focus about the Rasch model as it is important to examine the model level fit before looking at item level fit.
- *Item infit.* As illustrated in Section 2.4.1, item infit examines the fit of a particular item, and the items that do not fit the Rasch model can be identified and refined. We use a conventional acceptable range [0.7, 1.3].

5.4.2 Hyperparameter Settings

Table 5.2 shows the parameter values used for our parameter sensitivity experiments, where the default values in boldface are used in the experiments unless noted otherwise.

Table 5.2: Hyperparameter Settings

Parameter	Values Explored (**Default Value**)
λ_0	$[0.01, \mathbf{0.1}, 0.2, 0.3, 0.4, 0.5]$
λ_1	$[10^{-3}, 10^{-2}, \mathbf{10^{-1}}, 10^{0}, 10^{1}, 10^{2}]$
λ_2	$[10^{-3}, 10^{-2}, 10^{-1}, \mathbf{10^{0}}, 10^{1}, 10^{2}]$
λ_3	$[10^{-3}, 10^{-2}, 10^{-1}, \mathbf{10^{0}}, 10^{1}, 10^{2}]$
k	$[5, \mathbf{10}, 15, 20, 25, 30]$

5.4.3 Main Results for GG-NMF and TopicResponse

In this group of experiments, we examine the performance of GG-NMF (baseline) and TopicResponse in terms of negative log-likelihood, the quality of factorisation $\mathbf{WH}$ to $\mathbf{V}$, $\|\mathbf{V} - \mathbf{WH}\|_F^2$ and constraint $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$. For GG-NMF, the factorisation and Rasch estimation are separated, where topic-student response matrix $\mathbf{H}$ is firstly obtained using GG-NMF, and then is taken as input to Rasch estimation. For TopicResponse, the negative log-likelihood is minimised together with optimisation of factorisation, and can be obtained directly. The parameters are set using the values in boldface in Table 5.2. Figure 5.2 displays the negative log-likelihood of GG-NMF and TopicResponse.

It can be seen that TopicResponse can yield superior negative log-likelihood, implying better fit between the topic-student response matrix $\mathbf{H}$ and the Rasch model. TopicResponse therefore provides greater confidence that other item-level fit statistics such as infit, will be acceptable. *Jointly optimising the matrix factorisation and Rasch estimation can bring us closer to global optima.*

We present the results on quality of approximation $\|\mathbf{V} - \mathbf{WH}\|$ and supervision term $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|$, in Figure 5.3. From these plots, we can see that without sacrificing approximation performance in terms of $\|\mathbf{V} - \mathbf{WH}\|$, TopicResponse obtains superior $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|$ (while obtaining excellent negative

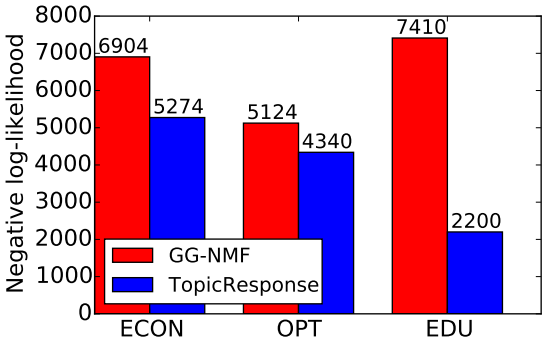
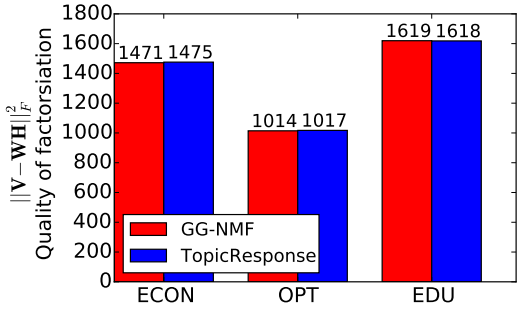
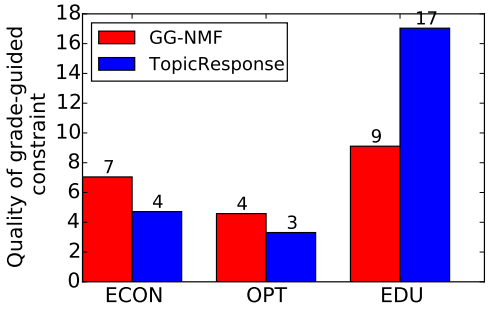


Figure 5.2: Negative log-likelihood of GG-NMF and TopicResponse; Smaller is better.



(a) Quality of factorisation, $\|\mathbf{V} - \mathbf{WH}\|_F^2$



(b) Quality of graded-guided constraint, $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$

Figure 5.3: Performance of GG-NMF and TopicResponse in terms of $\|\mathbf{V} - \mathbf{WH}\|_F^2$ and $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$; Smaller is better.

log-likelihoods as above). This performance again demonstrates that optimising the factorisation and Rasch estimation globally can be superior to optimising them separately. We therefore conclude that TopicResponse is preferable to GG-NMF; we focus on results for TopicResponse in the remainder of our experiments.

5.4.4 Item Infit, Item Difficulty and Student Ability

We further examine the infit of each item, which indicates if the set of topics conform to the Rasch model, and is appropriate for measurement. As illustrated in Section 2.4.1, a conventional acceptable range of infit is 0.7 to 1.3. As an

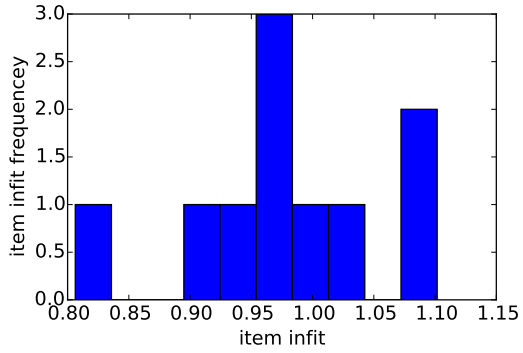


Figure 5.4: Item infit histogram for OPT MOOC

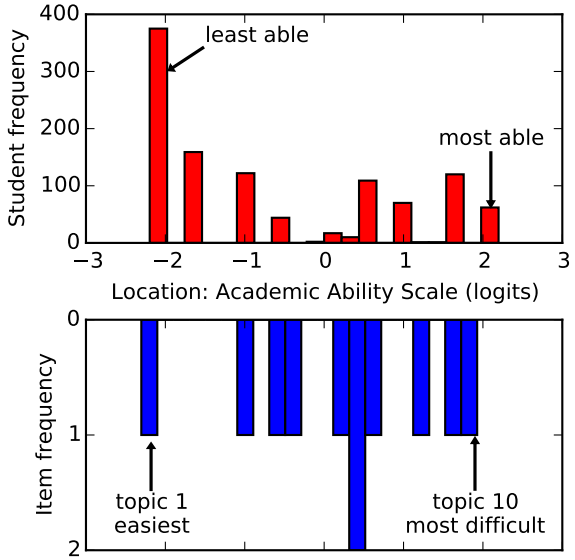


Figure 5.5: Histograms of OPT MOOC student ability location (top) and item difficulty location (bottom) on Rasch scale

example, we show the item infit in Figure 5.4 on OPT MOOC. We can see that the infit of each item is in the acceptable range, with most very close to the (ideal) expected value of 1.0, indicating that the set of topics conform to the Rasch model and is appropriate for measuring student ability.

Additionally, we examine item difficulties and student abilities. Figure 5.5 displays the histogram of item difficulty and student ability along a common scale. According to the Rasch model, the higher a person's ability relative to the difficulty of a topic, the higher the probability that person posts on that topic. It

can be seen that most students with low ability (around -2 logits), only dominate the "easiest" topic (topic 1 with difficult -2.3 logits); this topic concerns general problem solving. In other words, these students are likely to post only on topic 1, and unlikely to post on other topics. By comparison, the most able students with abilities around 2, with high probability contribute to all the topics.

5.4.5 Topic Interpretation and Discussion

We qualitatively examine topic interpretation, in order to assess educational meaningfulness. Well-scaled topics can potentially be used for curriculum refinement. Table 5.3 presents the topics generated using TopicResponse, alongside inferred difficulties. Topics are interpreted by an instructor who teaches a similar course. As the topics are not all course content-related, we envision that instructors examine discovered topics prior to using all for refining curriculum or taking other actions. Additionally, the inferred student ability levels and topic difficulty levels could be potentially used for personalised feedback, by tailoring appropriate topics of course content or forum discussion to students with their individual ability level taken into account. For example, most students (lowest ability) only discuss solving problem in general, as shown in Figure 5.5. If they cannot obtain sufficient help from forum discussions, they may be prone to drop out without further topic exploration. Therefore, in intervening with at-risk students, it is advisable to leverage discovered topics to better focus measures. Such services may be useful in preventing dropout in early stages (when most dropouts typically occur).

5.4.6 Parameter Sensitivity

To validate the robustness of TopicResponse to parameter settings, a series of sensitivity experiments were conducted. The parameter settings are shown in Table 5.2. Negative log-likelihoods, $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|$, $\|\mathbf{V} - \mathbf{WH}\|$ and $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|$ are examined in these experiments. Results for parameters λ_0, λ_1,λ_2,λ_3, and k on all three MOOCs are shown in Figure 5.6 to Figure 5.10.

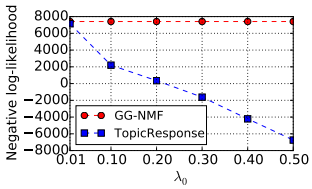
a) Effect of Parameter λ_0: As we can see from Figure 5.6 that as λ_0 increases,

Table 5.3: OPT MOOC topics generated from TopicResponse with inferred difficulty.

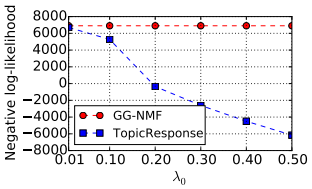
No.	Topics	Interpretation	Inferred difficulty
1	use time problem get solut one optim algorithm tri work	Solving in general	-2.30
2	cours thank would lectur realli great assign good like think	Course feedback	-0.93
3	python use run program solver java matlab instal command work	Python/Java/Matlab (How to start)	-0.63
4	problem thank solut get grade knapsack got feedback optim solv	How knapscak problem is solved and graded	-0.44
5	memori dp use column bb implement solv algorithm bound tabl	Comparing algorithms memory/time	0.23
6	color node graph random edg greedi opt search swap iter	Graph coloring	0.31
7	item valu weight capac estim take solut calcul best list	Knapsack problem	0.33
8	file pi line solver data submit lib urllib2 solveit open	Using solvers	0.52
9	video http class load lecture org problem coursera optimization 001	Platform	1.17
10	submit assign assignment error messag view assignment_id detail class coursera	Assignment submission	1.73

TopicResponse performs better in terms of negative log-likelihood, and performs worse in terms of the other three metrics due to the regularisation on the Rasch model, while the performance of GG-NMF does not change because there is no regularisation term on Rasch estimation. Overall, TopicResponse performs well when λ_0 varies between 0.1 and 0.2.

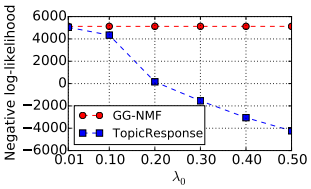
b) Effect of Parameter λ_1: As we can see from Figure 5.7, GG-NMF and TopicResponse are not sensitive to λ_1, performing stably with varying λ_1. TopicResponse constantly performs better in terms of negative log-likelihood while maintaining the comparable performance in terms of the other three metrics.



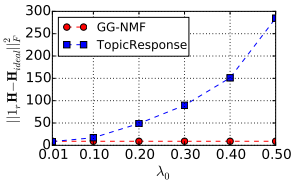
(a) Negative log-likelihood on EDU



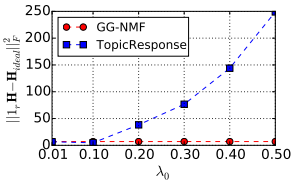
(b) Negative log-likelihood on ECON



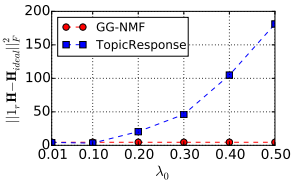
(c) Negative log-likelihood on OPT



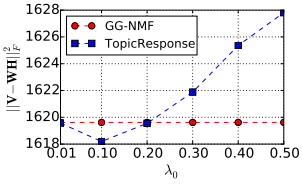
(d) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on EDU



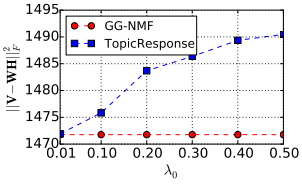
(e) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on ECON



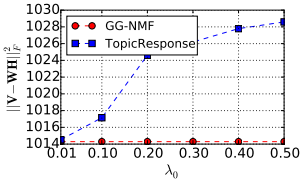
(f) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on OPT



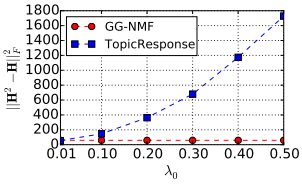
(g) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU



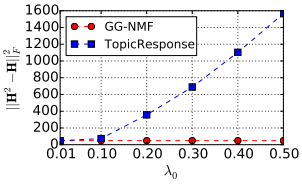
(h) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON



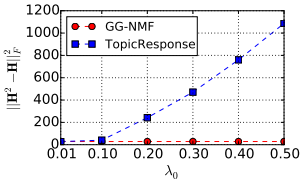
(i) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



(j) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on EDU



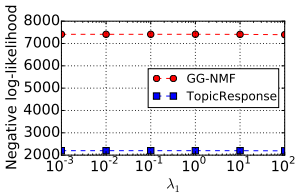
(k) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on ECON



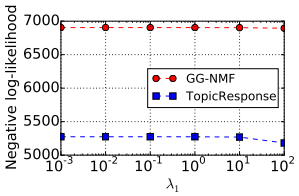
(l) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on OPT

Figure 5.6: Performance of GG-NMF and TopicResponse with varying λ_0.

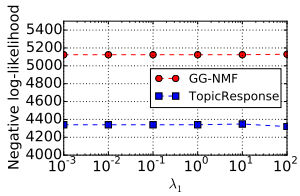
c) Effect of Parameter λ_2: It can be seen from Figure 5.8 that GG-NMF and TopicResponse perform well in terms of $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ (Figure 5.8d to Figure 5.8f) and $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ (Figure 5.8j to Figure 5.8l) when λ_2 varies from 10^0 to 10^2, and from 10^{-3} to 10^0 respectively. $\|\mathbf{V} - \mathbf{WH}\|_F^2$ gets worse as λ_1 increases, but does not change a lot compared to $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ and $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$. As λ_2 increases, the performance of GG-NMF and TopicResponse in terms of negative log-likelihood decrease, and TopicResponse constantly performs better than GG-NMF. Overall, λ_2 with values around 1.0 is good for GG-NMF and



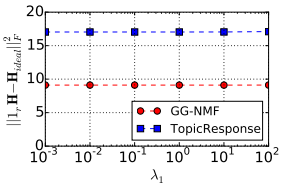
(a) Negative log-likelihood on EDU



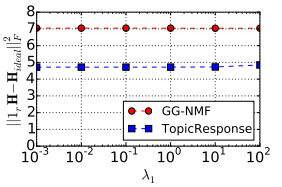
(b) Negative log-likelihood on ECON



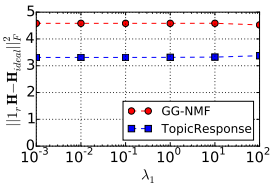
(c) Negative log-likelihood on OPT



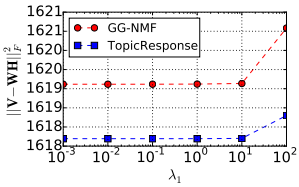
(d) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on EDU



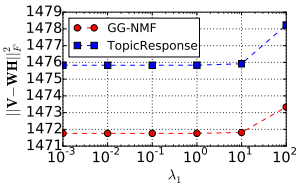
(e) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on ECON



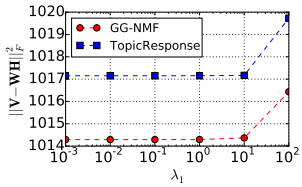
(f) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on OPT



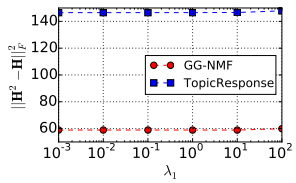
(g) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU



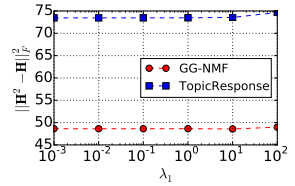
(h) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON



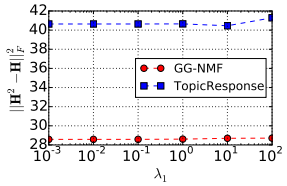
(i) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



(j) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on EDU



(k) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on ECON

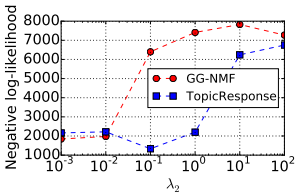


(l) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on OPT

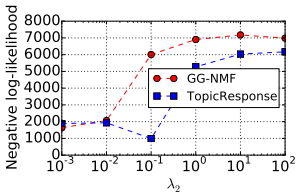
Figure 5.7: Performance of GG-NMF and TopicResponse with varying λ_1.

TopicResponse.

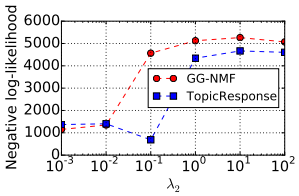
d) Effect of Parameter λ_3: It can be seen that GG-NMF and TopicResponse perform well in terms of $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ (Figure 5.9d to Figure 5.9f) and $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ (Figure 5.9j to Figure 5.9l) when λ_3 varies from 10^{-1} to 10^0, and from 10^0 to 10^2 respectively. Similar to λ_2, λ_3 does not affect $\|\mathbf{V} - \mathbf{WH}\|_F^2$ significantly. TopicResponse constantly achieves better negative log-likelihood than GG-NMF. Overall, λ_3 with values around 1.0 is good for GG-NMF and TopicResponse.



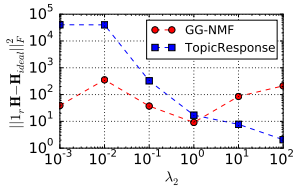
(a) Negative log-likelihood on EDU



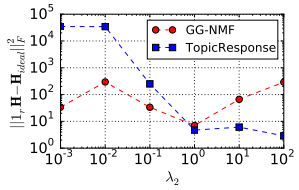
(b) Negative log-likelihood on ECON



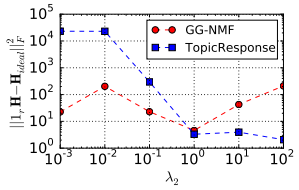
(c) Negative log-likelihood on OPT



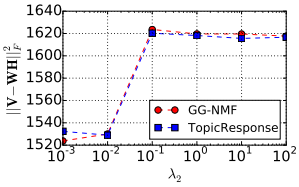
(d) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on EDU



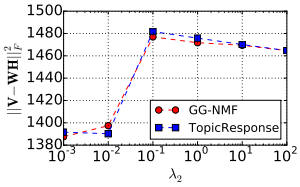
(e) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on ECON



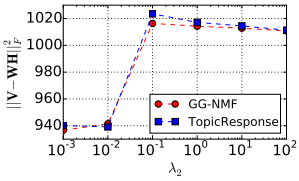
(f) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on OPT



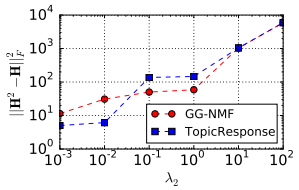
(g) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU



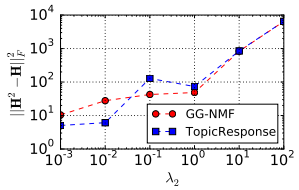
(h) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON



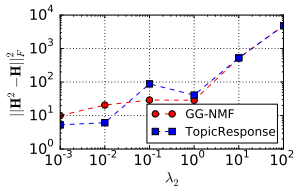
(i) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



(j) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on EDU



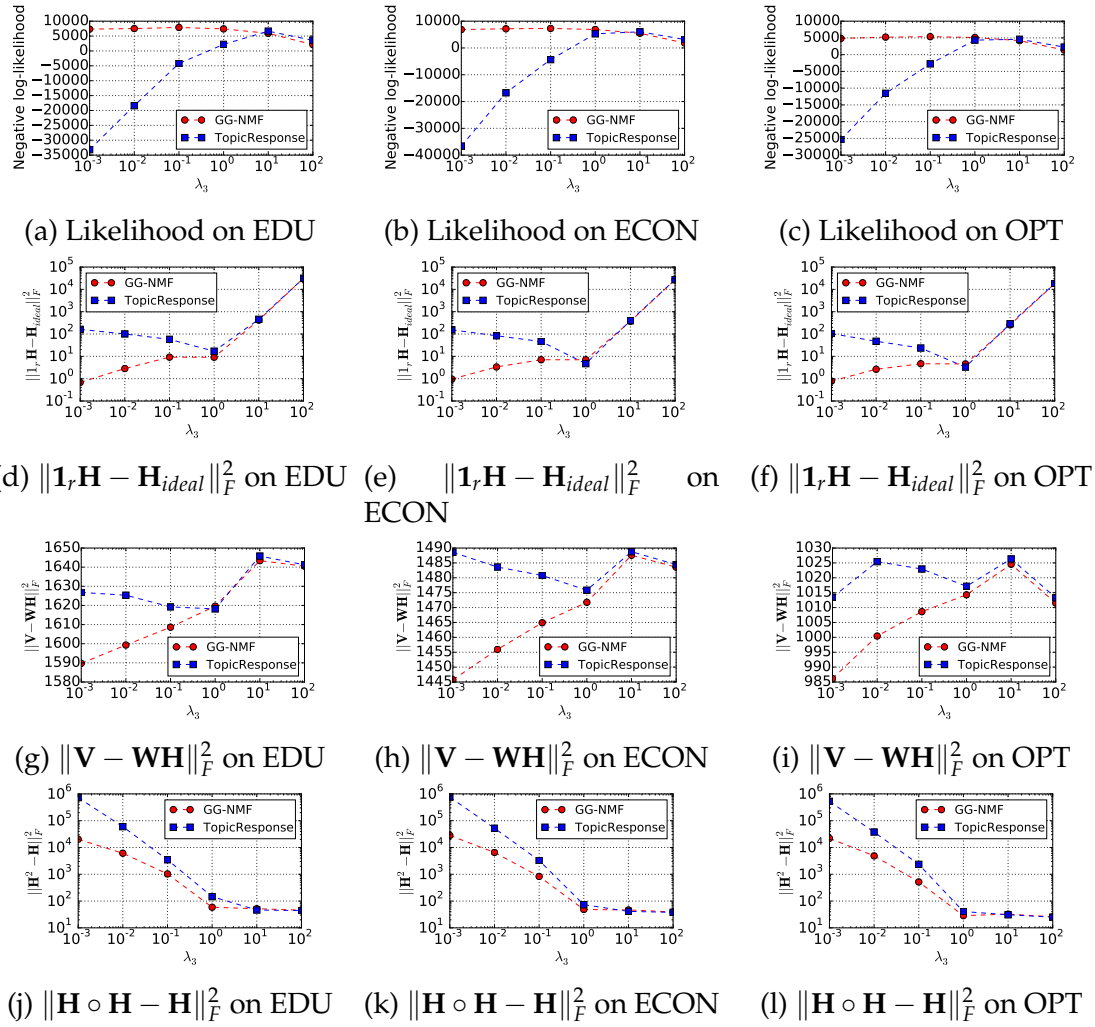
(k) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on ECON



(l) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on OPT

Figure 5.8: Performance of GG-NMF and TopicResponse with varying λ_2.

e) Effect of the number of topics k: It can be seen from Figure 5.10 that TopicResponse constantly outperforms GG-NMF in terms of negative log-likelihood, while getting slightly worse performance in the other three metrics. This is reasonable, as GG-NMF has more constraints and hence the model itself is less likely to perform as well as the less constrained GG-NMF in other metrics. Overall, GG-NMF and TopicResponse perform well in the experiments when k is set to 10 or 15. We choose 10 as the value of k since a smaller number of topics are easier to analyse.



(a) Likelihood on EDU (b) Likelihood on ECON (c) Likelihood on OPT

(d) $\|\mathbf{1}_r\mathbf{H}-\mathbf{H}_{ideal}\|_F^2$ on EDU (e) $\|\mathbf{1}_r\mathbf{H}-\mathbf{H}_{ideal}\|_F^2$ on ECON (f) $\|\mathbf{1}_r\mathbf{H}-\mathbf{H}_{ideal}\|_F^2$ on OPT

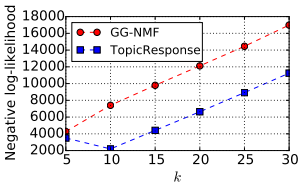
(g) $\|\mathbf{V}-\mathbf{WH}\|_F^2$ on EDU (h) $\|\mathbf{V}-\mathbf{WH}\|_F^2$ on ECON (i) $\|\mathbf{V}-\mathbf{WH}\|_F^2$ on OPT

(j) $\|\mathbf{H}\circ\mathbf{H}-\mathbf{H}\|_F^2$ on EDU (k) $\|\mathbf{H}\circ\mathbf{H}-\mathbf{H}\|_F^2$ on ECON (l) $\|\mathbf{H}\circ\mathbf{H}-\mathbf{H}\|_F^2$ on OPT

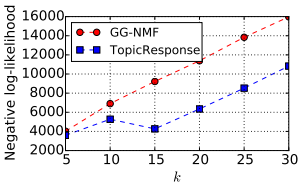
Figure 5.9: Performance of GG-NMF and TopicResponse with varying λ_3.

5.5 Conclusion

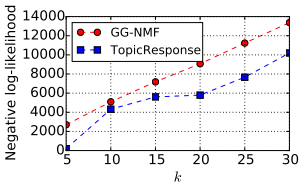
We have examined the suitability of content-based items (topics) discovered from MOOC forum discussions, for modelling student abilities. Our central tenet is that topics can be regarded as useful items for measuring latent skills, if student responses to these topics fit the Rasch item-response theory model, and if the discovered topics are further interpretable to domain experts. We propose to jointly optimise NMF and Rasch modelling, in order to discover Rasch-scaled topics. We provide a quantitative validation on three Coursera



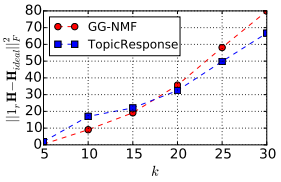
(a) Negative log-likelihood on EDU



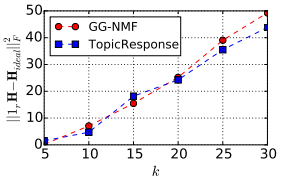
(b) Negative log-likelihood on ECON



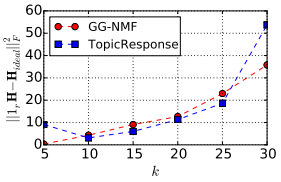
(c) Negative log-likelihood on OPT



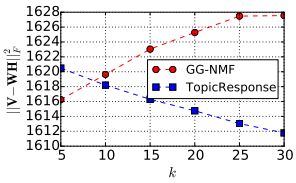
(d) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on EDU



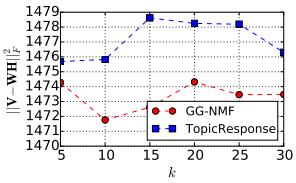
(e) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on ECON



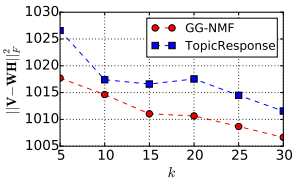
(f) $\|\mathbf{1}_r\mathbf{H} - \mathbf{H}_{ideal}\|_F^2$ on OPT



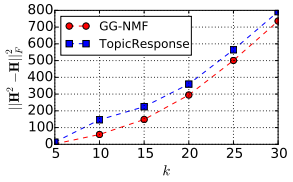
(g) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on EDU



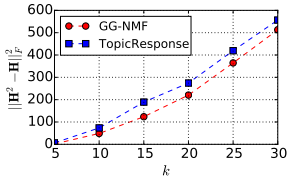
(h) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on ECON



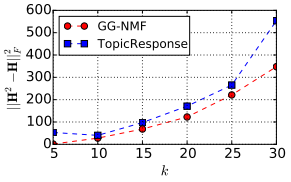
(i) $\|\mathbf{V} - \mathbf{WH}\|_F^2$ on OPT



(j) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on EDU



(k) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on ECON



(l) $\|\mathbf{H} \circ \mathbf{H} - \mathbf{H}\|_F^2$ on OPT

Figure 5.10: Performance of GG-NMF and TopicResponse with varying k.

MOOCs, demonstrating that TopicResponse yields better global fit to the Rasch model (observed with lower negative log-likelihood), maintains good quality of factorisation approximation, while measuring the students' academic abilities (reflected by the grade-guided constraint on students' participation on topics). We also provide qualitative examination of topic interpretation with inferred difficulty levels on a Discrete Optimisation MOOC. The results on goodness of fit and our qualitative examination, together suggest potential applications in curriculum refinement, student assessment and personalised feedback.

We opted to study the relatively simple Rasch model, as it forms the basis

of very many subsequent models in the literature. One direction for extension, is that for any model (like Rasch), that fits parameters via maximum-likelihood estimation (or risk minimisation in general), the model can be augmented with NMF as an additional regularisation. For example, such an extension should be straightforward for polychotomous observations, hierarchical models on latent skills, models that include more flexible per-student variation, etc. These represent fruitful direction for future research. Another possible extension could involve augmenting the $\mathbf{W}, \mathbf{H}$ matrices in the NMF or Rasch objective terms with manually-crafted items, to make effective use of prior knowledge.

Chapter 6

Similarity Search in Heterogeneous Information Networks

In this chapter, we study a generic problem of similarity search in heterogeneous information networks. We extend meta path-based similarity measure *PathSim* by incorporating transitive similarity and temporal dynamics. MOOCs contain diverse and complex student engagement data, which is a typical example of heterogeneous information networks, can benefit from this proposed similarity measure, *e.g.*, for finding similar students or threads, or thread recommendation by modelling student interactions in MOOC forums as a heterogeneous information network. This chapter is based on the following publication: Jiazhen He, James Bailey, and Rui Zhang. "Exploiting Transitive Similarity and Temporal Dynamics for Similarity Search in Heterogeneous Information Networks". In *International Conference on Database Systems for Advanced Applications*, pages 141–155. Springer, 2014.

6.1 Introduction

Heterogeneous information networks are ubiquitous in many real-world applications, such as bibliographic networks and healthcare networks. Different from homogeneous information networks (which only consider one type of object and link), heterogeneous information networks involve multiple types of

objects and links. For example, heterogeneous bibliographic networks contain authors as well as other types of objects, such as papers, venues, and terms. In addition, heterogeneous information networks contain rich semantic information. For example, two objects can be connected through different links with different semantic meanings (*i.e.*, two authors can be connected by co-authoring a paper or publishing different papers on a same venue). Such networks can more accurately model complex network data.

Heterogeneous information networks have been studied in many data mining tasks (Ji et al., 2011; Sun et al., 2009, 2010). In this chapter, we focus on the problem of similarity search in these networks. Similarity search aims to discover the most relevant objects with respect to a given query object. In heterogeneous information networks where multiple types of objects are available, we focus on identifying similar objects of the same type considering rich semantic information. For example, in a heterogeneous bibliographic network, given a query author, we can discover similar authors based on the diversified semantic meanings, such as co-author relationships and venues of publication.

Intuitively, two objects are similar if there are many paths between them. A major challenge for similarity search in heterogeneous information networks is how to exploit the diversified semantic meanings under different paths. Existing similarity measures for *homogeneous information networks* cannot effectively capture such meanings since they treat all the paths between two objects equally without distinguishing the different semantic meanings. Some existing studies have recognised this problem and tackled similarity search in *heterogeneous information networks* based on the concept of meta paths (Yu et al., 2012b; Sun et al., 2011b). A *meta path* is a sequence of links between object types, which can capture a particular semantic meaning between its starting type and ending type. The meta path-based similarity measures treat the concrete paths following a given meta path equally. However, the impacts of the paths connected through different objects can vary. The challenge is how to model such impacts. In addition, heterogeneous information networks evolve over time, and contain rich temporal information. For example, the link between two objects is generally formed with a timestamp. The challenge is how to exploit this temporal information for similarity search.

In this chapter, we extend the meta path-based similarity measure *PathSim* (Sun et al., 2011b) by incorporating transitive similarity and temporal information. A meta path can be concatenated by multiple short meta paths. Given a meta path, we first decompose it into multiple short meta paths with the start type and end type of the same type. For example, meta path "*author-paper-author-paper-author* " $(APAPA)$ describing two authors share same co-authors can be decomposed into two meta paths APA and APA. Then we add weights to the paths following a short meta path, according to the similarity between the two end objects of the short meta path, which is called transitive similarity. The transitive similarity between two objects can be obtained based on the different meta paths between them with different semantic meanings. The higher the transitive similarity between two objects, the more important the paths between them. For example, suppose two end authors x and y of $APAPA$ are connected through two common co-authors z_1 and z_2, if z_1 is more similar to x and y compared with z_2, the paths between x and z_1, and the ones between y and z_1 should be more important.

In addition, the paths between two objects are generally associated with temporal information, *i.e.*, the building time. Intuitively, the recent paths should be more important than old ones. The paths are generally built as a result of an event. For example, the path "$Tom - P_1 - SIGKDD$" with building time 2012 following the meta path "*author-paper-venuer*" is built due to the event that *Tom* published paper P_1 in $SIGKDD$ in 2012. To differentiate the importance of different paths, we first decompose a meta path into multiple short meta paths with the maximum length that an event can affect, for example, meta path "*author-paper-venue-paper-author* " can be decomposed into "*author-paper-venuer*" and "*venue-paper-author*". Then we add weights to the paths following the short meta paths according to their building time.

On the other hand, evaluating a new similarity measure is difficult, since it is difficult to obtain ground truth. We approach this challenge by assuming that similar objects will exhibit their similarity by their future behaviour. For example, in the Flickr image network, similar images are more likely to share the same tags or be in the same categories in the future. In bibliographic networks, similar authors are more likely to have collaborations in the future. Under this

assumption, we can obtain a ground truth to evaluate our extended similarity measure and compare it against existing methods.

The contributions of this chapter are summarised as follows:

- We develop a new method that incorporates transitive similarity to capture the impacts of different paths between two objects given a meta path.
- We incorporate temporal information for similarity search in heterogeneous information networks, by assigning different weights for the paths with different building time.
- Experiments on DBLP network data demonstrate the effectiveness of our proposed methods.

6.2 Related Work

An essential component for similarity search is a similarity measure, which measures the similarity between two objects. Similarity measures for traditional data types have been widely studied, for example the Jaccard coefficient and cosine similarity. For graph data, a number of studies utilise link information to measure the similarity between two objects. Early similarity measures include co-citation (Small, 1973) and co-coupling (Kessler, 1963), which were developed for scientific papers. Other similarity measures based on random walks have also been developed, such as SimRank (Jeh and Widom, 2002) and Personalised PageRank (Jeh and Widom, 2003). SimRank measures the similarity between two objects recursively, by averaging the similarity of their neighbours. Personalised PageRank measures the similarity between two objects by the probability of a random walk with restart starting from source object to target object.

The similarity measures defined in homogeneous networks ignore the different types of semantic information that is available under different paths in heterogeneous networks. There are several works on similarity search in heterogeneous information networks. Sun et al. (2011b) propose a meta path framework for heterogeneous information networks, where a meta path corresponds to a sequence of links between the objects. Based on the framework, a similarity

measure called *PathSim* is proposed, which aims to find similar objects with the same type. Yu et al. (2012b) study the problem of similarity query ambiguity, arising from the diversified semantic meanings in heterogeneous information networks. For a query object, users can provide example similar objects for the query as guidance for choosing related objects. Shi et al. (2012) study the problem of relevance search in heterogeneous networks, and propose a relevance measure called *HeteSim*, to measure the relatedness of the objects in heterogeneous networks, either of the same or different type. Overall, these works are based on the meta path framework and can capture semantic information under a meta path. However, they do not differentiate the impacts of concrete paths given a meta path, which can affect the similarity between two objects.

Another line of work related to our problem is link prediction, as the similarity between two objects can be used to predict the existence of a link between them (*i.e.*, friendships and co-authorship). In addition, since we evaluate the similarity measures considering the future behaviour between two similar objects, and such behaviour can be that a link will be formed between them in the future, our problem is similar to link prediction. However, we focus on developing similarity measures and the future information is only used for evaluation, while link prediction aims at developing methods to predict the existence of a link between two objects. The methods for link prediction can directly use similarity measures (Liben-Nowell and Kleinberg, 2003) or more sophisticated such as supervised learning (Al Hasan et al., 2006).

There are several works on link prediction in heterogeneous information networks (Sun et al., 2011a; Yu et al., 2012a; Sun et al., 2012). The most related to our problem is co-author relationship prediction in heterogeneous networks. Sun et al. (2011a), considering heterogeneous meta path-based features, use a logistic regression-based co-author relationship prediction model, to predict future co-author relationships. Our similarity measure can actually serve as a heterogeneous feature for their link prediction model.

6.3 Preliminaries and Problem Statement

In this section, we briefly introduce concepts related to heterogeneous information networks and define our problem.

A ***Heterogeneous information network*** is defined as a graph $G = (V, E, \mathcal{T}, \mathcal{R})$ where V is a set of objects, E is a set of links, $\mathcal{T}$ is a set of object types and $\mathcal{R}$ is a set of link types between object types. Since a heterogeneous information network contains multiple types of objects and links, $|\mathcal{T}| > 1$ and $|\mathcal{R}| > 1$. Each object $v \in V$ is associated with a particular type $T_i \in \mathcal{T}$, and each link $e \in E$ is associated with a particular type $R_j \in \mathcal{R}$.

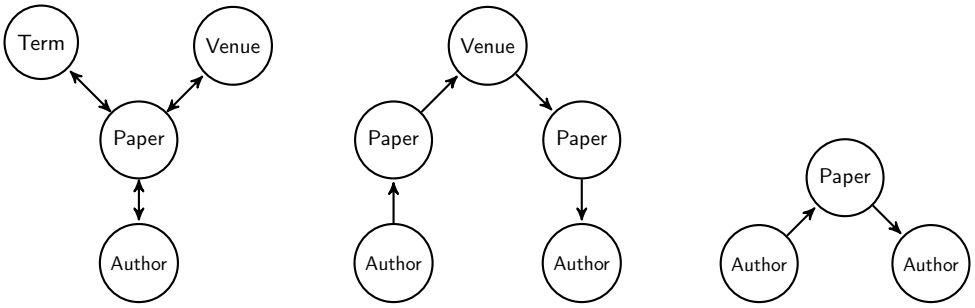
The concept of ***network schema*** (Sun et al., 2011b) has been proposed to describe the meta structure of a heterogeneous network for better understanding. It is a graph defined as $S_G = (\mathcal{T}, \mathcal{R})$ where each object is an object type and each link is a link type between object types.

For example, Figure 6.1a shows the network schema for a bibliographic information network. There are four types of objects: papers (P), venues (conferences/journals) (C), authors (A) and terms (T) which are the words appearing in the paper title. Also there are different links between the objects. For example, the links between authors and papers denote the writing or written-by relations.

A ***meta path*** $\mathcal{P}$ is a path defined over network schema, and is formalised as $T_1 \xrightarrow{R_1} T_2 \xrightarrow{R_2} \cdots \xrightarrow{R_l} T_{l+1}$, which defines a composite relation between type T_1 and T_{l+1}. The length of $\mathcal{P}$ is the number of relations in it. The objects can be connected through different meta paths. Two examples of meta path are shown in Figure 6.1b and Figure 6.1c. For simplicity, the meta path is denoted by the names of object types.

PathSim (Sun et al., 2011b) is a meta path-based similarity measure, which aims at finding similar peer objects for a query object, such as finding similar authors in terms of research area and reputation. Given a symmetric meta path $\mathcal{P}$, *PathSim* computes the similarity between two objects x and y according to

$$s(x, y) = \frac{2 \times |\mathcal{P}_{x \rightsquigarrow y}|}{|\mathcal{P}_{x \rightsquigarrow x}| + |\mathcal{P}_{y \rightsquigarrow y}|} \tag{6.1}$$



(a) Network schema (b) Meta path: APCPA (c) Meta path: APA

Figure 6.1: (a) A bibliographic network schema; (b) meta path "author-paper-venue-paper-author" (APCPA) describing authors publish papers in the same conferences; (c) meta path "author-paper-author" (APA) describing co-author relationship.

where $\mathcal{P}_{x \rightsquigarrow y}$ is the set of paths between x and y following $\mathcal{P}$, $\mathcal{P}_{x \rightsquigarrow x}$ is that between x and x, and $\mathcal{P}_{y \rightsquigarrow y}$ is that between y and y. The intuition behind *PathSim* is that two similar peer objects should not only be strongly connected, but also share comparable visibility. Their connectivity is defined as the number of paths between them following $\mathcal{P}$, and the visibility is defined as the number of paths between themselves (Sun et al., 2011b).

Given a symmetric meta path $\mathcal{P} = T_1 T_2 \cdots T_l$, *PathSim* similarity between two objects $x_i \in T_1$ and $x_j \in T_l$ with the same type $s(x_i, x_j)$, can be computed through the ***commuting matrix*** M, which is defined as $M = W_{T_1 T_2} W_{T_2 T_3} \cdots W_{T_{l-1} T_l}$, where $W_{T_i T_j}$ is the adjacency matrix between type T_i and type T_j. M_{ij} denotes the number of paths between object $x_i \in T_1$ and objects $y_j \in T_l$ following meta path $\mathcal{P}$, and $M_{ij} = |\mathcal{P}_{x_i \rightsquigarrow x_j}|$. Similarly, $M_{ii} = |\mathcal{P}_{x_i \rightsquigarrow x_i}|$ and $M_{jj} = |\mathcal{P}_{x_j \rightsquigarrow x_j}|$.

Problem Statement: The problem studied in this chapter is as follows. Given a heterogeneous information network and a query object, the goal is to find the top-k objects with the same type and the highest similarity with respect to the query object.

6.4 Proposed Methods

In this section, we present our methods which extend *PathSim* by incorporating transitive similarity and temporal information.

6.4.1 Transitive Similarity

Given a meta path $\mathcal{P} = T_1 T_2 \cdots T_l$, where T_1 and T_l are the same type ($T_1 = T_l$), $\mathcal{T}_m$ is the set of intermediate types which are the same as T_1 and T_l, $\mathcal{T}_m = (T_{m1}, T_{m2}, \cdots, T_{md})$ where d is the cardinality of $\mathcal{T}_m$. Therefore, $\mathcal{P}$ can be concatenated by multiple meta paths $\mathcal{P}_i (i = 1, \cdots, d+1)$, which is shown in Eq.(6.2).

$$\mathcal{P} = \underbrace{T_1 \cdots T_{m1}}_{\mathcal{P}_1} \underbrace{\cdots T_{m2}}_{\mathcal{P}_2} \cdots \underbrace{T_{md} \cdots T_l}_{\mathcal{P}_{d+1}} \tag{6.2}$$

PathSim (Sun et al., 2011b) treats all the paths between object $x \in T_1$ and $y \in T_l$ connected through different transitive objects $z \in T_{mh}$ equally. However, intuitively, we are more likely to trust the paths betweens the objects which are more similar to each other. We can put different weights on the paths following $\mathcal{P}_i$ considering the transitive similarity between the start type and the end type of $\mathcal{P}_i$. A simple way of obtaining the transitive similarity is to utilize *PathSim* over different meta paths with different semantic meanings. Therefore, for meta path $\mathcal{P}$, its commuting matrix can be computed as

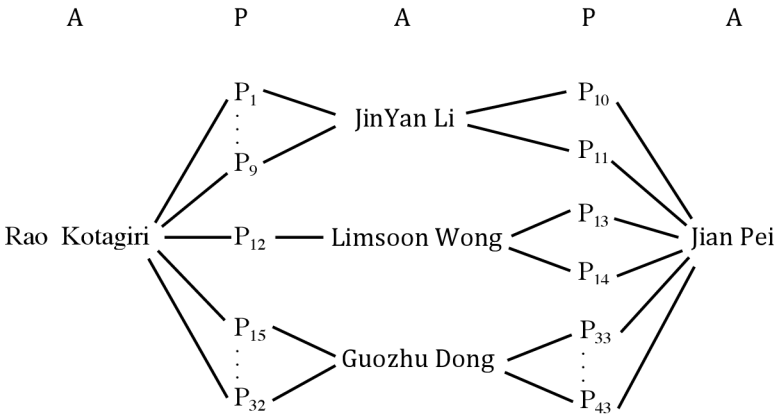
$$M_{\mathcal{P}} = M^s_{\mathcal{P}_1} M^s_{\mathcal{P}_2} \cdots M^s_{\mathcal{P}_{d+1}} \tag{6.3}$$

where $M^s_{\mathcal{P}_i}$ is the commuting matrix for meta path $\mathcal{P}_i$ with transitive similarity incorporated, and can be computed as

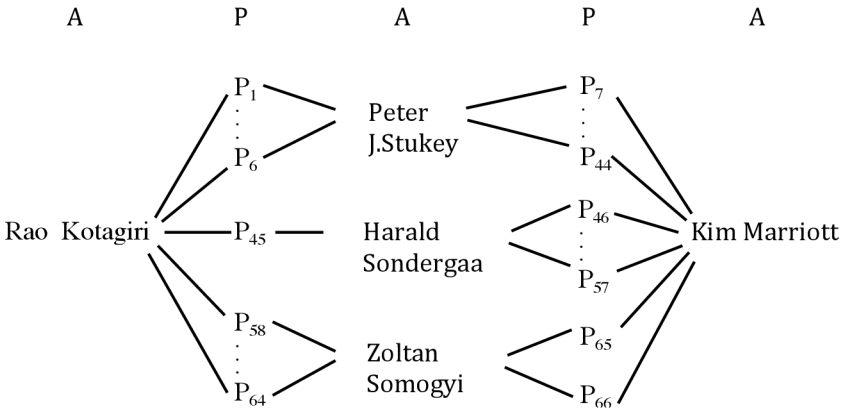
$$M^s_{\mathcal{P}_i} = M_{\mathcal{P}_i} \cdot S_{\mathcal{P}'} \tag{6.4}$$

where $M_{\mathcal{P}_i}$ denotes the commuting matrix of $\mathcal{P}_i$, with each element representing the number of paths between object $x \in T_s(\mathcal{P}_i)$ and object $y \in T_e(\mathcal{P}_i)$, where $T_s(\mathcal{P}_i)$ and $T_e(\mathcal{P}_i)$ represents the start type and the end type of $\mathcal{P}_i$ respectively.

$S_{\mathcal{P}'}$ denotes a transitive similarity matrix computed on meta path $\mathcal{P}'$. $\mathcal{P}'$ can be different meta paths such that $T_s(\mathcal{P}') = T_e(\mathcal{P}') = T_s(\mathcal{P}) = T_e(\mathcal{P})$. $S_{\mathcal{P}'}$ allows us to incorporate different meta paths with different semantic meanings.



(a) The paths between *Rao Kotagiri* and *Jian Pei* following *APAPA*



(b) The paths between *Rao Kotagiri* and *Kim Marriott* following *APAPA*

Figure 6.2: Example of paths following *APAPA* with *Rao Kotagiri* as the query author and two candidate authors

To better illustrate our method, we give an example in bibliographic networks. Figure 6.2 shows the paths between *Rao Kotagiri*(*Rao*) and *Jian Pei*(*Jian*)

following meta path $APAPA$, and the one between *Rao* and *Kim Marriott* (*Kim*) according to DBLP between 1990 and 2007. *Rao* and *Jian* (*Kim*) are not co-authors between 1990 and 2007. But they are connected through their common co-authors. Suppose *Rao* is the query author, the *PathSim* similarity between *Rao* and *Jian* according to Eq.(6.1) is,

$$s(Rao, Jian) = \frac{2 \times |APAPA_{Rao \rightsquigarrow Jian}|}{|APAPA_{Rao \rightsquigarrow Rao}| + |APAPA_{Jian \rightsquigarrow Jian}|} = \frac{2 \times (9 \times 2 + 1 \times 2 + 18 \times 11)}{21280 + 15333} = 0.0119$$

where the process of computation of $|APAPA_{Rao \rightsquigarrow Rao}| = 21280$ is not shown due to the space limitation, and the same for *Jian* (15333). Similarly, $s(Rao, Kim) = 0.0134$. However, according to our improved similarity measure,

$$s'(Rao, Jian) = \frac{2 \times \sum_{c \in Co}(|APA_{Rao \rightsquigarrow c}| \times S_{Rao,c} + |APA_{c \rightsquigarrow Jian}| \times S_{c,Jian})}{19357.04 + 12594.43} = \frac{2 \times 3.59}{19357.04 + 12594.43} = 2.25E - 04$$

where c denotes a common co-author of *Rao* and *Jian*, $Co = \{JinYan\ Li, Limsoon\ Wong, Guozhu\ Dong\}$ denotes the set of common co-authors of *Rao* and *Jian*, $S_{Rao,c}$ denotes the transitive similarity between *Rao* and c (in this example, S is computed based on APA), and similarly for $S_{c,Jian}$. The number of paths (weighted) between *Rao* and *Rao* (19357.04) is given directly due to the space limitation, and the same for *Jian* (12594.43). Similarly, $s'(Rao, Kim) = 1.43E - 04$. We assume that more similar authors are more likely to collaborate with the query author in future. In this example, based on the DBLP data between 2008 and 2013, *Jian* has collaboration with *Rao*, while *Kim* does not. We can see that our improved similarity measure can rank *Jian* higher compared with *Kim*.

6.4.2 Temporal Dynamics

Heterogenous information networks evolve over time, and also the similarity between two objects can change over time. We are more interested in finding similar objects now or even in the future. Intuitively, two objects are more

similar if there are more recent connections between them. Instead of treating the paths given a single snapshot equally, we differentiate the impacts of paths formed at different timestamps. A simple way is to put different weights on the paths formed in different timestamps. Essentially, the older paths make less contribution to similarity than recent ones, and should be given lower weights.

Given a meta path $\mathcal{P} = T_1 T_2 \cdots T_l$, its commuting matrix can be computed as

$$M_{\mathcal{P}} = M^t_{\mathcal{P}_1} M^t_{\mathcal{P}_2} \cdots M^t_{\mathcal{P}_g} \tag{6.5}$$

where $M^t_{\mathcal{P}_i}$ is the commuting matrix for meta path $\mathcal{P}_i$ with temporal information incorporated, and such that $\sum_{i=1}^{g} l(P_i) = l(\mathcal{P})$, where $l(\mathcal{P}_i)$ is the length of meta path $\mathcal{P}_i$. $\mathcal{P}_i$ is a meta path on which an event happens in a particular timestamp. For example, it can be APC in bibliographic networks which represents author publish paper in conference in a particular year. $M^t_{\mathcal{P}_i}$ can be computed as

$$M^t_{\mathcal{P}_i} = M_{\mathcal{P}_i} \cdot Y_{\mathcal{P}_i} \tag{6.6}$$

where $Y_{\mathcal{P}_i}$ is the temporal matrix on $\mathcal{P}_i$, with each element represents the weight of the path between object $x \in T_s(\mathcal{P}_i)$ and object $y \in T_e(\mathcal{P}_i)$. The weight can be assigned according to the timestamp of the path formed. Here, we define a function $f(t)$ of timestamp t to decide the weights,

$$f(t) = \alpha^{(t_1 - t)} (t_0 \leq t \leq t_1) \tag{6.7}$$

where t_0 and t_1 represent the start time and end time of the data used for computing similarities. $\alpha (0 < \alpha < 1)$ can be varied. The path formed most recently in t_1 has the largest weight 1. The smaller α is, the more rapidly the weight of the less recent path drops. Different $f(t)$ can be defined. We focus on the importance of incorporating temporal information instead of studying the impacts of different $f(t)$.

Based on the above proposed methods, we can improve *PathSim* by incorporating transitive similarity and/or temporal dynamic, and find the top-k similar objects for a give query object based on our improved similarity measure.

6.5 Experiments

In this section, we compare the effectiveness of our improved similarity measure using the *PathSim* measure as a baseline.

6.5.1 Evaluation Measure

Assessing similarity is challenging since it is difficult to obtain ground truth providing a quantitative measure for the similarity between two objects. Most existing methods to evaluate the performance of similarity measures rely on user studies or on an reliable external measure of similarity. The study in Sun et al. (2011b) used case studies and manually labeled the results for a handful of queries. In this work, since we assume that similar objects will show similar behaviour in some way in the future, we can obtain ground truth to evaluate the similarity measure and provide a comprehensive experimental assessment using thousands of test queries.

We use NDCG (Discounted Normalised Cumulative Gain), a widely used measure in information retrieval (Agichtein et al., 2006; Balasubramanian et al., 2010), to evaluate the ranking performance. It rewards relevant objects in the top ranked results more heavily than those ranked lower. In particular, we use NDCG@n, which computes NDCG over the top n ranked objects, and which can be computed as

$$\begin{aligned} NDCG@n &= \frac{DCG@n}{IDCG@n} \\ DCG@n &= rel(1) + \sum_{i=2}^{n} \frac{rel(x_i)}{log_2(i)} \end{aligned} \tag{6.8}$$

where $IDCG@n$ denotes the Ideal DCG for a perfect ranking and $rel(x_i)$ denotes the relevance score for an object x_i at position i.

6.5.2 Experiment Setup

The DBLP dataset downloaded on 25th April 2013 is used in our experiments. The network schema of DBLP network is same as Figure 6.1a. The data from

1990 to 2007 (denoted as $T_{1990-2007}$) is used to compute similarity, while the data from 2008 to 2013 (denoted as $T_{2008-2013}$) is used for evaluation. The number of authors, papers, conferences (including journals) and terms (after removing stopwords in paper titles) between 1990 and 2007 are shown in Table. 6.1.

Table 6.1: DBLP data between 1990 and 2007

Data	Author	Paper	Conference	Term
1990-2007	698,507	1,114,726	4,949	139,613

We focus on computing the similarity between two authors given a meta path between them. In particular, we use meta path $APAPA$ which implies two authors share the same co-authors. Given a query author q, the top n similar authors are returned with similarity computed based on the data in $T_{1990-2007}$. We assume that similar authors will exhibit their similarity by their future behaviour. For meta path $APAPA$, two similar authors might collaborate in the future ($T_{2008-2013}$). To easily capture such behaviour for evaluation, we only return the top n similar authors who have not collaborated with the query author in $T_{1990-2007}$. To evaluate the ranking performance, we need the relevance score $rel(x_i)$ for each returned similar author w.r.t. q. According to the number of co-authored publications between x_i and q in $T_{2008-2013}$, $rel(x_i)$ can be set as

$$rel(x_i) = \begin{cases} 0 & \text{if } N(q, x_i)\text{=0} \\ \varphi\left(N(q, x_i)\right) & \text{if } N(q, x_i) \neq 0 \end{cases} \tag{6.9}$$

where $N(q, x_i)$ denotes the number of papers that q and x_i publish together in $T_{2008-2013}$. We use $\mathcal{C}$ to denote the set of all the candidate authors. The candidate authors are ranked in ascending order according to $N(q, x)(x \in \mathcal{C})$, and each candidate is assigned a ranking value according to its ranking position. For those who have same value of $N(q, x)$, the same ranking value will be assigned. $\varphi(\cdot)$ is a mapping function from $N(q, x_i)$ to the ranking value for x_i.

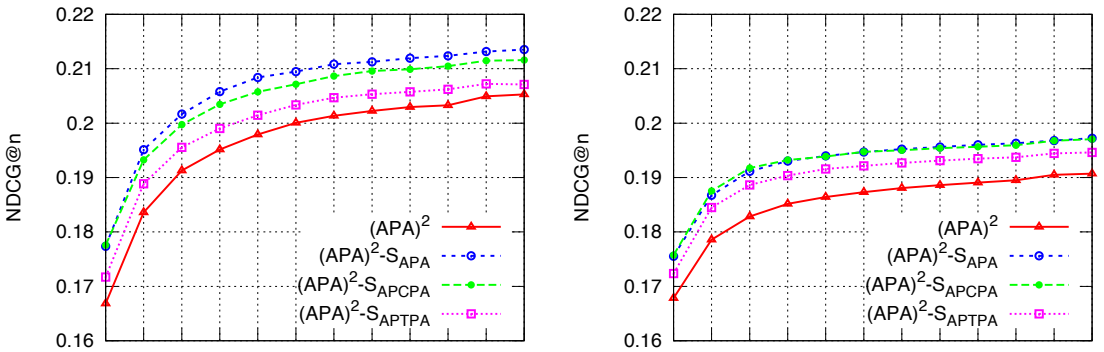
The query authors can be chosen from the set of authors who exist in $T_{1990-2007}$, and have new collaborations with authors exist in $T_{1990-2007}$ in future time interval $T_{2008-2013}$. We randomly select 3000 authors as query authors, and compute

the averaged results over the 3000 authors. We compare our improved similarity measure with *PathSim* using paired t-test with $p = 0.05$. This process is repeated 10 times, and the results reported in this chapter are the averaged results over 10 runs. In addition, we show the effectiveness of our similarity measure on two sets of query authors, highly productive authors with more than 15 publications in $T_{1990-2007}$ (denoted as HP), and less productive authors with between 5 and 15 publications in $T_{1990-2007}$ (denoted as LP).

6.5.3 Experimental Results

Transitive Similarity Incorporated.

In this group of experiments, we incorporate different kinds of transitive similarity into meta path $APAPA$. We compare our methods with the baseline method, *PathSim* applied on $APAPA$. The results are shown in Figure 6.3, where $(APA)^2$ represents the baseline method, and $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ represents our methods on $APAPA$ with incorporated transitive similarity based on APA, $APCPA$ and $APTPA$ respectively. All the results have statistical significance with p-value$<$0.05.



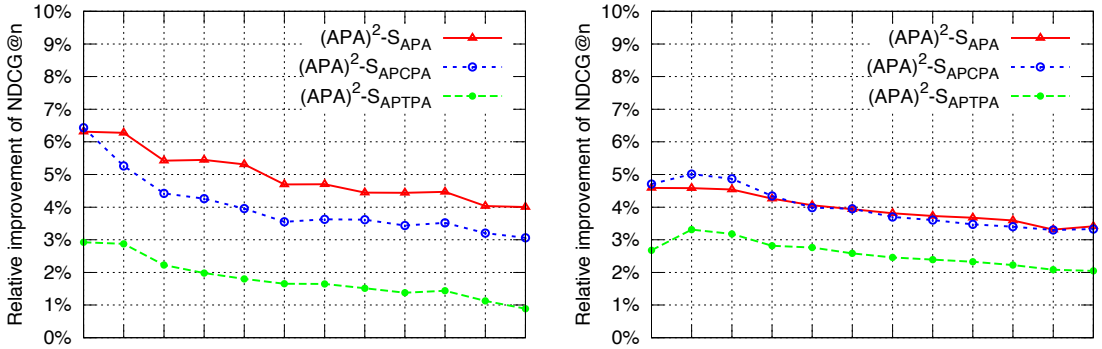
(a) HP queries

(b) LP queries

Figure 6.3: NDCG@n of $(APA)^2$ denoting the baseline method (*PathSim*) on $APAPA$ and our methods $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ denoting $APAPA$ with incorporated transitive similarity based on APA, $APCPA$ and $APTPA$ respectively, for (a)HP queries and (b)LP queries.

It can be seen from Figure 6.3 that after incorporating different similarity information, the performances of our methods are improved over all the varying n on both HP and LP queries. Essentially, the similarity incorporated based on APA gives better performance compared with $APCPA$ and $APTPA$. In addition, the performances of all the similarity measures in terms of NDCG@N are low. The main reason is that ranking is generally difficult, especially in the case of similar authors in terms of future collaborators, and only using the raw similarity produced by the similarity measures. Actually, two authors can collaborate due to many external factors that cannot be captured using the similarity measures in this chapter. Another reason is that for each run, among the 3000 queries, there are a number of queries with 0 for NDCG@n , which degrade the average results. Such queries do not have future collaborations with their 2-hop authors.

In addition, the overall performance of both the baseline method and our methods on LP queries is worse than that on HP queries. The reason is that for each run, among the 3000 queries, only about 1500 queries have new collaborations with their 2-hop authors for LP queries, while about 2200 for HP queries. Meanwhile, it indicates that HP authors are more likely to collaborate with their 2-hop authors compared with LP authors.



(a) HP queries

(b) LP queries

Figure 6.4: Relative improvements of our methods $(APA)^2 - S_{APA}$, $(APA)^2 - S_{APCPA}$ and $(APA)^2 - S_{APTPA}$ over $PathSim$ on $APAPA$

Since the absolute improvements can be misleading, we mainly report the

relative improvements of NDCG@n (which is also used in studies in information retrieval (Qin et al., 2007; Yeh et al., 2007)) in the following experiments. The relative improvements of our methods over *PathSim* on meta path $APAPA$ are given in Figure 6.4. We can see that the relative improvements of our method with transitive similarity S_{APA} and S_{APCPA}, are more than 4% and 3% respectively over all the values of varying n on HP queries. Furthermore, the relative improvements for S_{APTPA} on HP queries is less than that on LP queries. The reason might be that HP authors are generally active in diverse research topics, which yields diverse terms.

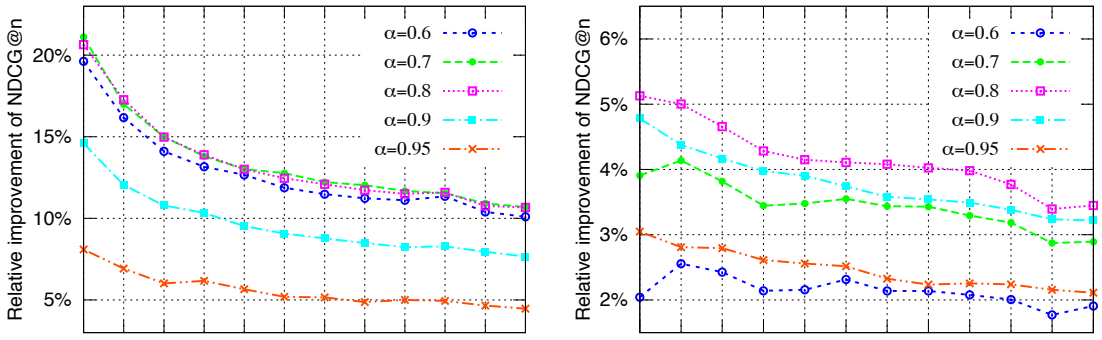
Temporal Information Incorporated.

In this group of experiments, we show the effectiveness of incorporating temporal information. We incorporate temporal information into meta path $APAPA$, and use Eq.(6.7) to decide the weights of the paths following APA. Here, $t_0 = 1990$, $t_1 = 2007$.

First we study the impact of parameter α. Figure 6.5 shows the relative improvements of our method $(APA)^2\_T_\alpha$ with varying α over *PathSim* on $APAPA$, where $(APA)^2\_T_\alpha$ denotes incorporating the temporal information (with varying α) into $APAPA$. It can be seen that when $\alpha = 0.8$, our method can yield good performance on both HP and LP queries. In addition, the relative improvements on HP queries are much higher than LP queries. The reason might be that the links associated with LP authors are relatively sparse, and are formed in a relatively short time interval, which do not contain much diversified temporal information to be exploited.

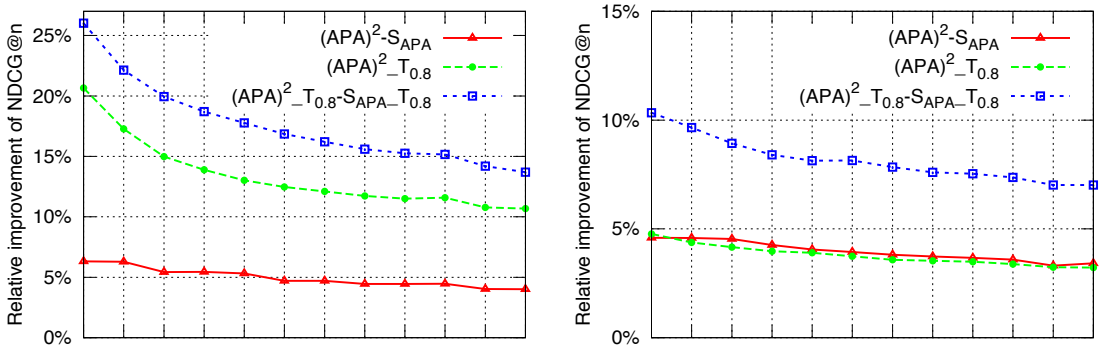
Furthermore, we compare the relative improvements over *PathSim* when incorporating temporal information and/or transitive similarity into $APAPA$. Figure 6.6 shows the results when incorporating only transitive similarity ($(APA)^2\_S_{APA}$), only temporal information ($(APA)^2\_T_{0.8}$), and both of them ($APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$) to $APAPA$.

It can be seen that there is little difference for the relative improvements of incorporating transitive similarity on HP queries and LP queries. But incorporating temporal information makes huge differences, and basically it works bet-



(a) HP queries (b) LP queries

Figure 6.5: Relative improvements of our method $(APA)^2\_T_\alpha$ denoting the temporal information (with varying α) incorporated to $APAPA$ over $PathSim$ on $APAPA$.



(a) HP queries (b) LP queries

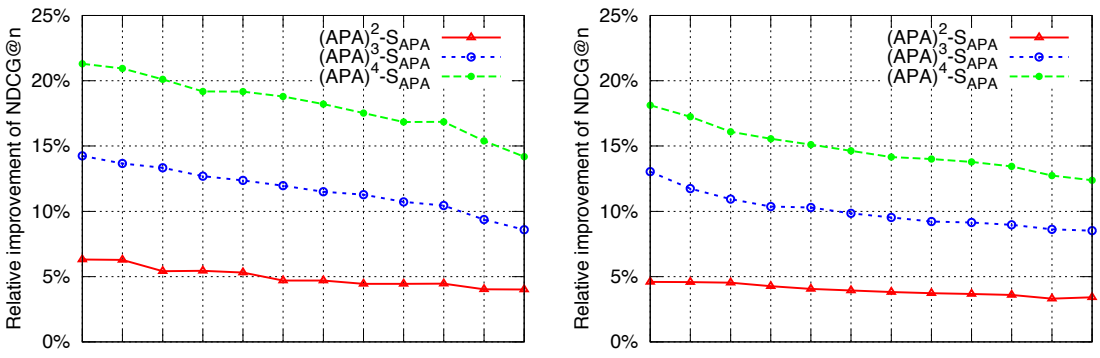
Figure 6.6: Relative improvements of our method $(APA)^2\_S_{APA}$, $(APA)^2\_T_{0.8}$ and $APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$ over $PathSim$ on HP queries and LP queries.

ter for HP queries. In addition, the more information incorporated, the higher the performance is, which can be seen from Figure 6.6 that, $APAPA\_T_{0.8} - S_{APA}\_T_{0.8}$ achieves the best performance with relative improvements more than 15% on HP queries and more than 7% on LP queries.

Impacts on Different Length of Meta Path

In this group of experiments, we check the impacts of transitive similarity on different length of meta path. Figure 6.7 shows the relative improvements of incorporating transitive similarity (based on APA) into different length of meta path APA over $PathSim$ applied on corresponding length of meta path APA, where $(APA)^4 - S_{APA}$ represents the relative improvements of incorporating transitive similarity (based on APA) into $(APA)^4$ over $PathSim$ on $(APA)^4$, and similarly for $(APA)^3 - S_{APA}$ and $(APA)^2 - S_{APA}$.

It can be seen that the relative improvement on longer paths is much higher than shorter paths. This is because $PathSim$ does not distinguish the importance of different paths given a meta path. When increasing the length of a meta path, $PathSim$ will treat more remote (and possibly irrelevant) neighbours as similar, whilst our methods which take into account transitive similarity can alleviate this effect.



(a) HP queries

(b) LP queries

Figure 6.7: Relative improvement on NDCG@n for different length of APA with transitive similarity based on APA incorporated

6.6 Conclusion

We study the problem of similarity search in heterogeneous information networks and propose an improved meta path-based similarity measure which in-

corporates transitive similarity and temporal information. Experimental results in DBLP networks show that our improved similarity measures outperforms the baseline existing method for identifying similar authors for future collaboration. We also found that using temporal information can provide greater gains on highly productive authors than less productive authors. Furthermore, using transitive similarity and temporal information simultaneously can produce the best performance. Although the similarity measure has been proposed and evaluated in the domain of DBLP bibliographic network, it could readily apply to MOOC settings to find similar students or threads, or thread recommendation, by modelling the student interactions in MOOCs forums as a heterogeneous information network.

Chapter 7

Conclusion and Future Work

7.1 Summary of Contributions

In this thesis, we focused on applying and developing machine learning algorithms for the analysis of MOOC data to assist instructors in providing feedback to students in order to improve learning experiences and learning outcomes. In particular, we focused on three themes: a) identifying students who are at risk of not completing courses for timely intervention; b) exploring the suitability of using automatically discovered forum topics as instruments to measure student ability; c) similarity search in heterogeneous information networks.

Chapter 3 studied the first theme. We built prediction models week by week. Predicted probabilities served a dual purpose, both for the identification of at-risk students and as the basis for subsequent intervention. We suggested an intervention that presented the probability of success/failure to borderline students. This intervention required smoothed and well-calibrated probabilities. Based on regularised logistic regression, we proposed two transfer learning algorithms, LR-SEQ and LR-SIM, to balance accuracy with smoothness. Experiments on Coursera MOOCs showed the effectiveness of both LR-SEQ and LR-SIM. LR-SIM, leveraging knowledge in later weeks performed better in early weeks than LR-SEQ, which only uses previous week's knowledge, which is desirable for early intervention. We also showed that the prediction models trained on a first offering work well on a second offering. The early identifica-

tion of at-risk students could help instructors design interventions, and the suggested intervention could assist instructors in providing feedback to borderline students.

Chapters 4 and 5 studied the second theme but considered two different measurement models as evidence of reliable measurement, the Guttman scale and the Rasch model respectively. This is a novel problem that has not been studied previously, but is of importance for both gaining insight into student learning and our algorithm contribution lies in combining machine learning model and measurement model. Chapter 4 discovered Guttman-scaled topics by adding a regularisation term to NMF to enforce that students' participation across topics conforms to the Guttman scale. Quantitative experiments on three Coursera MOOCs established the statistical effectiveness of our algorithm. Furthermore, we provided a qualitative result giving domain experts' interpretation on two MOOCs, which supported the understandability and inferred difficulty ranking of the scaled topics. The results suggested their potential applicability in curriculum refinement and student assessment.

Continuing the theme of Chapter 4, Chapter 5 explored the same problem, but used a different measurement model, Rasch model from Item Response Theory for examining the statistical effectiveness of reliability. This model is more widely used in education and psychology. The Gutttman scale only allows ordering of students and topics in a latent scale, while the Rasch model makes the distance between them meaningful. Therefore, it could be potentially used for adaptive assessment and providing personalised feedback. We have proposed to jointly optimise NMF and Rasch modelling, in order to discover Rasch-scale topics. We provided quantitative validation on three Coursera MOOCs, and a qualitative examination of topic interpretation on a Discrete Optimisation MOOC. The results on statistical effectiveness and qualitative examination suggested potential applications in curriculum refinement, student assessment and personalised feedback.

Chapter 6 studied the third theme of finding similar objects in heterogeneous information networks, where rich semantic information are available. We extended the meta path-based similarity measure *PathSim* by incorporating richer information, such as transitive similarity and temporal dynamics, and evalu-

ated on a large DBLP bibliographic network. Results showed that our improved similarity measure is more effective at identifying similar authors in terms of their future collaborations. The proposed similarity measure could apply to MOOC settings for finding similar students or threads, or thread recommendation, by modelling student interactions in MOOC forums as a heterogeneous information network.

7.2 Limitations and Future Work

7.2.1 Intervention for Non-Borderline Students

Chapter 3 suggested an intervention targeted at students who are in the pass/fail borderline, which can only potentially help a small fraction of students compared with the large number of students enrolled. It would be interesting to study the possible interventions for high-risk students and low-risk students. Existing studies have suggested the importance of collaborative learning between peers to improve student learning. Students' forum behaviours could be used together with their risk levels to form a study group. For example, high-risk students seeking help in forums and low-risk students actively providing help in forums could be potentially grouped together. More sophisticatedly, the topic difficulty levels inferred in Chapter 5 could be used to further match their needs. For example, low-risk students actively providing help could help high-risk students who are seeking help on topics with similar difficulty levels.

7.2.2 Detailed Feedback

Model interpretability is important in learning analytics, where detailed feedback may be favoured over generic feedback like "how's it going?". Such specifics can shed light on why a student is failing, and also what strategies other students follow to succeed. In particular, within logistic regression, the learned weight vectors can be used for explaining the contribution of each feature—albeit under certain assumptions on feature correlation. In these cases, features are not only important for prediction, but also for interpretability. It would be interest-

ing to study models which yields both strong prediction and good interpretability, *e.g.*, combining neural networks (good prediction) with pattern-based models (good interpretability).

7.2.3 Content-Based Measurement

The combination of topic models and measurement models for content-based measurement opens a number of exciting directions for further research. Broadly speaking, the consequences of content-based measurement on educational theories and practice requires further understanding, while the study of statistical models for psychometrics by computer science can stimulate interesting new machine learning.

Additionally, Chapters 4 and 5 only consider the use of forum content as instruments for student ability measurement. It would be interesting to examine the student engagement patterns (*e.g.*, watching videos, clickstream observations, completing assignments, posting behaviours in forums, etc.) together with forum content to devise a comprehensive set of items for measurement. This will make further use of the rich data generated in MOOCs, and benefit applications such as providing personalised and detailed feedback to students based on their ability levels and behaviour difficulty levels.

Alternatively, our approach could be extended to incorporate partial prior knowledge. For example, an education researcher or instructor might already possess certain items extracted from student engagement behaviours in MOOCs based on their domain knowledge. It would be useful to discover topics that can be used together with the existing items to measure the predefined latent attribute.

7.2.4 Personalised Recommendation

Item response theory enables the meaningful positioning of students and items on a latent scale, with student ability and item difficulty inferred. It would be interesting to see this applied in recommender systems to provide personalised recommendation. For example, in Chapter 5, with the student ability and topic

difficulty levels, one can recommend appropriate threads for students based on their individual ability. So the students won't be shown threads too easy or too difficult for them. Furthermore, algorithmically, it would be interesting to see how information about student ability and topic difficulty could be incorporated and combined with recommender techniques. For example, such information could be used as additional information or constraint in matrix factorisation for thread recommendation.

Bibliography

Hervé Abdi. Guttman scaling. In Neil J. Salkind, editor, *Encyclopedia of Research Design*. SAGE Publications, 2010.

Eugene Agichtein, Eric Brill, and Susan Dumais. Improving web search ranking by incorporating user behavior information. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 19–26. ACM, 2006.

Mohammad Al Hasan, Vineet Chaoji, Saeed Salem, and Mohammed Zaki. Link prediction using supervised learning. In *Proceedings of the Sixth SIAM Data Mining Workshop on Link Analysis, Counter-terrorism and Security*, 2006.

Ashton Anderson, Daniel Huttenlocher, Jon Kleinberg, and Jure Leskovec. Engaging with massive online courses. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 687–698. International World Wide Web Conferences Steering Committee, 2014.

Rie Kubota Ando and Tong Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *The Journal of Machine Learning Research*, 6:1817–1853, 2005.

Rel Guzman Apaza, Elizabeth Vera Cervantes, Laura Cruz Quispe, and José Ochoa Luna. Online courses recommendation based on LDA. In *1st Symposium on Information Management and Big Data*, page 42, 2014.

Thushari Atapattu and Katrina Falkner. A framework for topic generation and labeling from mooc discussions. In *Proceedings of the Third (2016) ACM Conference on Learning@ Scale*, pages 201–204. ACM, 2016.

Yoram Bachrach, Thore Graepel, Tom Minka, and John Guiver. How to grade a test without knowing the answers—a bayesian graphical model for adaptive crowdsourcing and aptitude testing. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 1183–1190, 2012.

Frank B Baker and Seock-Ho Kim. *Item response theory: Parameter estimation techniques*. CRC Press, 2004.

Girish Balakrishnan. Predicting student retention in massive open online courses using hidden markov models. Master's thesis, EECS Department, University of California, Berkeley, May 2013. URL `http://www.eecs.berkeley.edu/Pubs/TechRpts/2013/EECS-2013-109.html`.

Niranjan Balasubramanian, Giridhar Kumaran, and Vitor R Carvalho. Predicting query performance on the web. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 785–786. ACM, 2010.

Michael F Beaudoin. Learning or lurking?: Tracking the "invisible" online student. *The Internet and Higher Education*, 5(2):147–155, 2002.

Yoav Bergner, Stefan Droschler, Gerd Kortemeyer, Saif Rayyan, Daniel Seaton, and David E Pritchard. Model-based collaborative filtering analysis of student response data: Machine-learning item response theory. *International Educational Data Mining Society*, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Trevor G Bond and Christine M Fox. Applying the rasch model: Fundamental measurement in the human sciences. 2001.

Jordan L Boyd-Graber, David M Blei, and Xiaojin Zhu. A topic model for word sense disambiguation. In *EMNLP-CoNLL*, pages 1024–1033, 2007.

John Champaign, Kimberly F Colvin, Alwina Liu, Colin Fredericks, Daniel Seaton, and David E Pritchard. Correlating skill and improvement in 2

MOOCs with a student's time on tasks. In *Proceedings of the first ACM conference on Learning@ scale conference*, pages 11–20. ACM, 2014.

Snigdha Chaturvedi, Dan Goldwasser, and Hal Daumé III. Predicting instructor's intervention in MOOC forums. In *ACL (1)*, pages 1501–1511, 2014.

Amit Chauhan. Massive open online courses (MOOCs): Emerging trends in assessment and accreditation. *Digital Education Review*, (25):7–17, 2014.

Wang Chong, David Blei, and Fei-Fei Li. Simultaneous image classification and annotation. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1903–1910. IEEE, 2009.

Cody A Coleman, Daniel T Seaton, and Isaac Chuang. Probabilistic use cases: Discovering behavioral patterns for predicting certification. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 141–148. ACM, 2015.

Kimberly F Colvin, John Champaign, Alwina Liu, Colin Fredericks, and David E Pritchard. Comparing learning in a MOOC and a blended on-campus course. In *Educational Data Mining 2014*, 2014.

Rafael Jaime De Ayala. *Theory and practice of response theory*. Guilford Publications, 2013.

Jennifer DeBoer, Andrew Ho, Glenda S Stump, David E Pritchard, Daniel Seaton, and Lori Breslow. Bringing student backgrounds online: MOOC user demographics, site usage, and online learning. *engineer*, 2:0–81, 2013a.

Jennifer DeBoer, GS Stump, D Seaton, and Lori Breslow. Diversity in MOOC students' backgrounds and behaviors in relationship to performance in 6.002 x. In *Proceedings of the Sixth Learning International Networks Consortium Conference*, 2013b.

Laura Dietz, Steffen Bickel, and Tobias Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, pages 233–240. ACM, 2007.

Jorge Dıez, O Luaces, A Alonso-Betanzos, A Troncoso, and A Bahamonde. Peer assessment in MOOCs using preference learning via matrix factorization. In *NIPS Workshop on Data Driven Education*, 2013.

Li Fei-Fei and Pietro Perona. A bayesian hierarchical model for learning natural scene categories. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 2, pages 524–531. IEEE, 2005.

Sean Gerrish and David M Blei. A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pages 375–382, 2010.

Nabeel Gillani, Rebecca Eynon, Michael Osborne, Isis Hjorth, and Stephen Roberts. Communication communities in MOOCs. *arXiv preprint arXiv:1403.4640*, 2014.

L Guttman. The basis for scalogram analysis. In S. Stouffer, editor, *Measurement and Prediction: The American Soldier*. Wiley, New York, 1950.

Sherif Halawa, Daniel Greene, and John Mitchell. Dropout prediction in MOOCs using learner activity features. *Experiences and best practices in and around MOOCs*, page 7, 2014a.

Sherif Halawa, Daniel Greene, and John Mitchell. Dropout prediction in MOOCs using learner activity features. In *Proceedings of the European MOOC Summit*, 2014b.

Julie Hare. Melbourne university's MOOCs hit enrolment record. *The Australian*, 2016.

Jiazhen He, James Bailey, Benjamin IP Rubinstein, and Rui Zhang. Identifying at-risk students in massive open online courses. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 1749–1755. AAAI Press, 2015.

Jiazhen He, Benjamin IP Rubinstein, James Bailey, Rui Zhang, Sandra Milligan, and Jeffrey Chan. MOOCs meet measurement theory: A topic-modelling ap-

proach. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, pages 1195–1201. AAAI Press, 2016.

Diane Hu and Lawrence K Saul. A probabilistic topic model for unsupervised learning of musical key-profiles. In *ISMIR*, pages 441–446. Citeseer, 2009.

Glen Jeh and Jennifer Widom. Simrank: a measure of structural-context similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '02, pages 538–543, Edmonton, Alberta, Canada, 2002. ACM. ISBN 1-58113-567-X. doi: 10.1145/775047.775126.

Glen Jeh and Jennifer Widom. Scaling personalized web search. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 271–279, Budapest, Hungary, 2003. ACM. ISBN 1-58113-680-3. doi: 10.1145/775152.775191.

Maximilian Jenders, Ralf Krestel, and Felix Naumann. Which answer is best? predicting accepted answers in MOOC forums. *WWW'16 Companion*, 2016.

Ming Ji, Jiawei Han, and Marina Danilevsky. Ranking-based classification of heterogeneous information networks. In *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1298–1306. ACM, 2011.

Katy Jordan. Initial trends in enrolment and completion of massive open online courses. *The International Review of Research in Open and Distributed Learning*, 15(1), 2014.

Charlie Kaufman, Radia Perlman, and Mike Speciner. *Network Security: Private Communication in a Public World*. Prentice Hall, 2nd edition, 2002.

Maxwell Mirton Kessler. Bibliographic coupling between scientific papers. *American documentation*, 14(1):10–25, 1963.

Hanan Khalil and Martin Ebner. MOOCs completion rates and possible methods to improve retention-a literature review. In *World Conference on Educational Multimedia, Hypermedia and Telecommunications*, number 1, pages 1305–1313, 2014.

René F Kizilcec, Chris Piech, and Emily Schneider. Deconstructing disengagement: analyzing learner subpopulations in massive open online courses. In *Proceedings of the third international conference on learning analytics and knowledge*, pages 170–179. ACM, 2013.

Marius Kloft, Felix Stiehler, Zhilin Zheng, and Niels Pinkwart. Predicting MOOC dropout over weeks using machine learning methods. In *Proceedings of the EMNLP Workshop on Modeling Large Scale Social Interaction in Massively Open Online Courses*, 2014.

Steve Kolowich. Coursera takes a nuanced view of MOOC dropout rates. *The chronicle of higher education*, 2013.

Daniel D Lee and H Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

Daniel D Lee and H Sebastian Seung. Algorithms for non-negative matrix factorization. In *Advances in Neural Information Processing Systems*, pages 556–562, 2001.

David Liben-Nowell and Jon Kleinberg. The link prediction problem for social networks. In *Proceedings of the twelfth international conference on Information and knowledge management*, CIKM '03, pages 556–559, New Orleans, LA, USA, 2003. ACM. ISBN 1-58113-723-0. doi: 10.1145/956863.956972.

John M Linacre. What do infit and outfit, mean-square and standardized mean. *Rasch Measurement Transactions*, 16(2):878, 2002.

John M Linacre. Misfit diagnosis: Infit outfit mean-square standardized. *Retrieved June*, 1:2006, 2006.

J. M. Lodge. What if student attrition was treated like an illness? an epidemiological model for learning analytics. In G. Williams, P. Statham, N. Brown, and B. Cleland, editors, *Changing Demands, Changing Directions. Proceedings ascilite Hobart 2011*, pages 822–825, 2011.

Frederic M Lord. *Applications of item response theory to practical testing problems*. Routledge, 1980.

Jenny Mackness, Sui Mak, and Roy Williams. The ideals and reality of participating in a MOOC. 2010.

Fei Mi and Dit-Yan Yeung. Probabilistic graphical models for boosting cardinal and ordinal peer grading in MOOCs. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Sandra Milligan. Crowd-sourced learning in MOOCs: learning analytics meets measurement theory. In *Proceedings of the Fifth International Conference on Learning Analytics And Knowledge*, pages 151–155. ACM, 2015.

Alexandru Niculescu-Mizil and Rich Caruana. Obtaining calibrated probabilities from boosting. In *UAI*, page 413, 2005a.

Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine learning*, pages 625–632, 2005b.

Daniel FO Onah, Jane Sinclair, and Russell Boyatt. Dropout rates of massive open online courses: behavioural patterns. *EDULEARN14 Proceedings*, pages 5825–5834, 2014.

Laura Pappano. The year of the MOOC. *The New York Times*, 2(12):2012, 2012.

Elazar J Pedhazur and Liora Pedhazur Schmelkin. *Measurement, design, and analysis: An integrated approach (student ed.)*. Lawrence Erlbaum Associates, Inc, 1991.

Chris Piech, Jonathan Huang, Zhenghao Chen, Chuong Do, Andrew Ng, and Daphne Koller. Tuned models of peer assessment in MOOCs. *arXiv preprint arXiv:1307.2579*, 2013.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Tao Qin, Xu-Dong Zhang, De-Sheng Wang, Tie-Yan Liu, Wei Lai, and Hang Li. Ranking with multiple hyperplanes. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 279–286. ACM, 2007.

A Ramesh, D Goldwasser, B Huang, H Daume III, and L Getoor. Understanding MOOC discussion forums using seeded lda. In *Proc. of 9th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 28–33, 2014a.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daumé III, and Lise Getoor. Modeling learner engagement in MOOCs using probabilistic soft logic. In *NIPS Workshop on Data Driven Education*, 2013.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014b.

Arti Ramesh, Dan Goldwasser, Bert Huang, Hal Daume III, and Lise Getoor. Learning latent engagement patterns of students in online courses. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI Press, 2014c.

Arti Ramesh, Shachi H Kumar, James Foulds, and Lise Getoor. Weakly supervised models of aspect-sentiment for online course discussion forums. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

Georg Rasch. Probabilistic models for some intelligence and attainment tests. *Copenhagen: Danish Institute for Educational Research*, 1960.

Ry Rivard. Measuring the MOOC dropout rate. *Inside Higher Ed*, 8:2013, 2013.

Anthony C Robinson. Exploring class discussions from a massive open online course (MOOC) on cartography. In *Modern Trends in Cartography*, pages 173–182. Springer, 2015.

A Zand Scholten. Admissible statistics from a latent variable perspective. *Theory & Psychology*, 18:111–117, 2011.

Chuan Shi, Xiangnan Kong, Philip S. Yu, Sihong Xie, and Bin Wu. Relevance search in heterogeneous networks. In *Proceedings of the 15th International Conference on Extending Database Technology*, EDBT '12, pages 180–191, Berlin, Germany, 2012. ACM. ISBN 978-1-4503-0790-1. doi: 10.1145/2247596.2247618.

Henry Small. Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science*, 24(4):265–269, 1973.

Yizhou Sun, Yintao Yu, and Jiawei Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *Proceedings of the 15th ACM SIGKDD international Conference on Knowledge Discovery and Data Mining*, pages 797–806. ACM, 2009.

Yizhou Sun, Jie Tang, Jiawei Han, Manish Gupta, and Bo Zhao. Community evolution detection in dynamic heterogeneous information networks. In *Proceedings of the Eighth Workshop on Mining and Learning with Graphs*, pages 137–146. ACM, 2010.

Yizhou Sun, Rick Barber, Manish Gupta, Charu C. Aggarwal, and Jiawei Han. Co-author relationship prediction in heterogeneous bibliographic networks. In *Proceedings of the 2011 International Conference on Advances in Social Networks Analysis and Mining*, ASONAM '11, pages 121–128, Washington, DC, USA, 2011a. IEEE Computer Society. ISBN 978-0-7695-4375-8. doi: 10.1109/ASONAM.2011.112.

Yizhou Sun, Jiawei Han, Xifeng Yan, Philip S Yu, and Tianyi Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. *Proceedings of the VLDB Endowment*, 4(11), 2011b.

Yizhou Sun, Jiawei Han, Charu C. Aggarwal, and Nitesh V. Chawla. When will it happen?: relationship prediction in heterogeneous information networks. In *Proceedings of the fifth ACM international conference on Web search and data mining*, WSDM '12, pages 663–672, Seattle, Washington, USA, 2012. ACM. ISBN 978-1-4503-0747-5. doi: 10.1145/2124295.2124373.

Colin Taylor, Kalyan Veeramachaneni, and Una-May O'Reilly. Likely to stop? predicting stopout in massive open online courses. *arXiv preprint arXiv:1408.3382*, pages 273–275, 2014.

Warren S Torgerson. Theory and methods of scaling. 1958.

Kristina Toutanova and Mark Johnson. A bayesian lda-based model for semi-supervised part-of-speech tagging. In *Advances in neural information processing systems*, pages 1521–1528, 2007.

Kalyan Veeramachaneni, Una-May O'Reilly, and Colin Taylor. Towards feature engineering at scale for data from massive open online courses. *CoRR*, abs/1407.5238, 2014. URL `http://arxiv.org/abs/1407.5238`.

Xing Wei and W Bruce Croft. Lda-based document models for ad-hoc retrieval. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 178–185. ACM, 2006.

Miaomiao Wen and Carolyn Penstein Rosé. Identifying latent study habits by mining learner behavior patterns in massive open online courses. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, pages 1983–1986. ACM, 2014.

Miaomiao Wen, Diyi Yang, and Carolyn Penstein Rosé. Sentiment analysis in MOOC discussion forums: What does it tell us? *Proceedings of Educational Data Mining*, 2014.

Benjamin D Wright and Geofferey N Masters. *Rating Scale Analysis. Rasch Measurement.* ERIC, 1982.

Benjamin D Wright, John M Linacre, JE Gustafson, and P Martin-Lof. Reasonable mean-square fit values. *Rasch measurement transactions*, 8(3):370, 1994.

Bin Xu and Dan Yang. Study partners recommendation for xMOOCs learners. *Computational Intelligence and Neuroscience*, 2015, 2015.

Diyi Yang, Tanmay Sinha, David Adamson, and Carolyn Penstein Rosé. Turn on, tune in, drop out: Anticipating student dropouts in massive open online courses. In *Proceedings of the 2013 NIPS Data-Driven Education Workshop*, 2013.

Diyi Yang, David Adamson, and Carolyn Penstein Rosé. Question recommendation with constraints for massive open online courses. In *Proceedings of the 8th ACM Conference on Recommender systems*, pages 49–56. ACM, 2014a.

Diyi Yang, Miaomiao Wen, and Carolyn Rose. Peer influence on attrition in massive open online courses. *Proceedings of Educational Data Mining*, 2014b.

Diyi Yang, Miaomiao Wen, Iris Howley, Robert Kraut, and Carolyn Rose. Exploring the effect of confusion in discussion forums of massive open online courses. In *Proceedings of the Second (2015) ACM Conference on Learning@ Scale*, pages 121–130. ACM, 2015.

Jen-Yuan Yeh, Jung-Yi Lin, Hao-Ren Ke, and Wei-Pang Yang. Learning to rank for information retrieval using genetic programming. In *Proceedings of SIGIR 2007 Workshop on Learning to Rank for Information Retrieval (LR4IR 2007)*, 2007.

Xiao Yu, Quanquan Gu, Mianwei Zhou, and Jiawei Han. Citation prediction in heterogeneous bibliographic networks. In *Proceedings of the Twelfth SIAM International Conference on Data Mining*, pages 1119–1130, Anaheim, California, USA, 2012a.

Xiao Yu, Yizhou Sun, Brandon Norick, Tiancheng Mao, and Jiawei Han. User guided entity similarity search using meta-path selection in heterogeneous information networks. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, CIKM '12, pages 2025–2029, Maui, Hawaii, USA, 2012b. ACM. ISBN 978-1-4503-1156-4. doi: 10.1145/2396761.2398565.

Zhong-Yuan Zhang, Tao Li, Chris Ding, Xian-Wen Ren, and Xiang-Sun Zhang. Binary matrix factorization for analyzing gene expression data. *Data Mining and Knowledge Discovery*, 20(1):28–52, 2010.

Zhongyuan Zhang, Chris Ding, Tao Li, and Xiangsun Zhang. Binary matrix factorization with applications. In *Data Mining, 2007. ICDM 2007. Seventh IEEE International Conference on*, pages 391–400. IEEE, 2007.

Leon Wenliang Zhong and James T Kwok. Accurate probability calibration for multiple classifiers. In *Proceedings of the Twenty-Third international joint conference on Artificial Intelligence*, pages 1939–1945. AAAI Press, 2013.